Sumário

1 Conceitos Básicos **1**
1.1 Introdução . . . 1
1.2 Espaço Vetorial . . . 2
1.3 Processo de Gram-Schmidt . . . 15
1.4 Projeção Ortogonal . . . 18
1.5 Auto-Valores e Auto-Vetores . . . 22
1.6 Exercícios Complementares . . . 30

2 Análise de Arredondamento em Ponto Flutuante **32**
2.1 Introdução . . . 32
2.2 Sistema de Números Discreto no Computador . . . 32
2.3 Representação de Números no Sistema $F(\beta, t, m, M)$ . . . 37
2.4 Operações Aritméticas em Ponto Flutuante . . . 40
2.5 Efeitos Numéricos . . . 42
2.5.1 Cancelamento . . . 43
2.5.2 Propagação do erro . . . 44
2.5.3 Instabilidade Numérica . . . 46
2.5.4 Mal Condicionamento . . . 48
2.6 Exercícios Complementares . . . 51

3 Equações não Lineares **55**
3.1 Introdução . . . 55
3.2 Iteração Linear . . . 61
3.3 Método de Newton . . . 68
3.4 Método das Secantes . . . 71
3.5 Método Regula Falsi . . . 74
3.6 Sistemas de Equações não Lineares . . . 77
3.6.1 Iteração Linear . . . 77
3.6.2 Método de Newton . . . 79
3.7 Equações Polinomiais . . . 82
3.7.1 Determinação de Raízes Reais . . . 83
3.7.2 Determinação de Raízes Complexas . . . 86
3.7.3 Algoritmo Quociente-Diferença . . . 91
3.8 Exercícios Complementares . . . 94
3.9 Problemas Aplicados e Projetos . . . 97

4 Solução de Sistemas Lineares: Métodos Exatos **108**
4.1 Introdução . . . 108
4.2 Decomposição LU . . . 113
4.3 Método de Eliminação de Gauss . . . 119
4.4 Método de Gauss-Compacto . . . 127
4.5 Método de Cholesky . . . 130
4.6 Método de Eliminação de Gauss com Pivotamento Parcial . . . 135
4.7 Refinamento da Solução . . . 137
4.8 Mal Condicionamento . . . 140
4.9 Cálculo da Matriz Inversa . . . 143
4.10 Exercícios Complementares . . . 146
4.11 Problemas Aplicados e Projetos . . . 150

5 Solução de Sistemas Lineares: Métodos Iterativos **156**
5.1 Introdução . . . 156
5.2 Processos Estacionários. . . . 156
5.2.1 Método de Jacobi-Richardson . . . 158
5.2.2 Método de Gauss-Seidel. . . . 162
5.3 Processos de Relaxação . . . 168
5.3.1 Príncipios Básicos do Processo de Relaxação . . . 171
5.3.2 Método dos Gradientes . . . 173
5.3.3 Método dos Gradientes Conjugados . . . 176
5.4 Exercícios Complementares . . . 181
5.5 Problemas Aplicados e Projetos . . . 185

6 Programação Matemática **191**
6.1 Espaço Vetorial . . . 191

7 Determinação Numérica de Auto-Valores e Auto-Vetores **192**
7.1 Introdução . . . 192
7.2 Método de Leverrier . . . 194
7.3 Método de Leverrier-Faddeev . . . 195
7.4 Método das Potências . . . 201
7.4.1 Método da Potência Inversa . . . 205
7.4.2 Método das Potências *com Deslocamento* . . . 207
7.5 Auto-Valores de Matrizes Simétricas . . . 211
7.5.1 Método Clássico de Jacobi . . . 213
7.5.2 Método Cíclico de Jacobi . . . 218
7.6 Método de Rutishauser (ou Método LR) . . . 220
7.7 Método de Francis (ou Método QR) . . . 223
7.8 Exercícios Complementares . . . 227
7.9 Problemas Aplicados e Projetos . . . 230

8 Aproximação de Funções: Método dos Mínimos Quadrados **234**
8.1 Introdução . . . 234
8.2 Aproximação Polinomial . . . 235
8.2.1 Caso Contínuo . . . 235
8.2.2 Caso Discreto: . . . 241
8.2.3 Erro de Truncamento . . . 245
8.3 Aproximação Trigonométrica . . . 246

8.3.1 Caso Contínuo 246
8.3.2 Caso Discreto 250
8.4 Outros Tipos de Aproximação 252
8.5 Sistemas Lineares Incompatíveis 262
8.6 Exercícios Complementares 265
8.7 Problemas Aplicados e Projetos 268

9 Programação não Linear 279

10 Aproximação de Funções: Métodos de Interpolação Polinomial 280
10.1 Introdução 280
10.2 Polinômio de Interpolação 280
10.3 Fórmula de Lagrange 283
10.4 Erro na Interpolação 287
10.5 Interpolação Linear 290
10.6 Fórmula para Pontos Igualmente Espaçados 292
10.7 Outras Formas do Polinômio de Interpolação 296
10.7.1 Diferença Dividida 297
10.7.2 Cálculo Sistemático das Diferenças Divididas. 297
10.7.3 Alguns Resultados sobre Diferenças Divididas 299
10.7.4 Fórmula de Newton 299
10.7.5 Diferenças Ordinárias 306
10.7.6 Cálculo Sistemático das Diferenças Ordinárias 307
10.7.7 Fórmula de Newton-Gregory 310
10.8 Exercícios Complementares 312
10.9 Problemas Aplicados e Projetos 316

11 Integração Numérica 321
11.1 Introdução 321
11.2 Fórmulas de quadratura interpolatória 322
11.2.1 Fórmulas de Newton-Cotes 325
11.2.2 Erro nas Fórmulas de Newton-Cotes 335
11.3 Polinômios Ortogonais 340
11.3.1 Principais Polinômios Ortogonais 342
11.3.2 Propriedades dos Polinômios Ortogonais 344
11.4 Fórmulas de Quadratura de Gauss 348
11.4.1 Fórmula de Gauss-Legendre 351
11.4.2 Fórmula de Gauss-Tchebyshev 353
11.4.3 Fórmula de Gauss-Laguerre 355
11.4.4 Fórmula de Gauss-Hermite 356
11.5 Erro nas Fórmulas de Gauss 357
11.6 Exercícios Complementares 361
11.7 Problemas Aplicados e Projetos 366

12 Solução Numérica de Equações Diferenciais Ordinárias 379
12.1 Introdução 379
12.2 Método de Taylor de Ordem q 380
12.3 Métodos Lineares de Passo Múltiplo 383
12.3.1 Obtidos do Desenvolvimento de Taylor 384
12.3.2 Obtidos de Integração Numérica 386

12.3.3 Ordem e Constante do Erro . . . 389
12.3.4 Erro de Truncamento Local . . . 391
12.3.5 Consistência e Estabilidade . . . 393
12.3.6 Convergência . . . 394
12.4 Métodos do Tipo Previsor - Corretor . . . 397
12.4.1 Erro de Truncamento Local . . . 399
12.5 Método Geral Explícito de 1-passo . . . 401
12.5.1 Ordem . . . 401
12.5.2 Consistência . . . 401
12.5.3 Convergência . . . 402
12.5.4 Métodos de Runge-Kutta . . . 403
12.6 Sistemas de Equações e Equações de Ordem Elevada . . . 412
12.6.1 Sistemas de Equações Diferenciais . . . 413
12.6.2 Equações Diferenciais de Ordem Elevada . . . 418
12.7 Exercícios Complementares . . . 420
12.8 Problemas Aplicados e Projetos . . . 423

13 Solução Numérica de Equações Diferenciais Parciais 429
13.1 Introdução . . . 429
13.2 Equações Parabólicas . . . 431
13.3 Métodos de Diferenças Finitas . . . 436
13.4 Problemas Não Lineares . . . 453
13.5 Equações Parabólicas em Duas Dimensões . . . 455
13.6 Equações Elípticas . . . 460
13.7 Métodos de Diferenças Finitas . . . 462
13.8 Erro de Truncamento Local . . . 465
13.9 Condições de Fronteira em Domínios Gerais . . . 469
13.10 Condição de Fronteria de Neumann . . . 473
13.11 Diferenças Finitas em Coordenadas Polares . . . 475
13.12 Exercícios . . . 476

14 Exercícios Mistos 486

Capítulo 1

Conceitos Básicos

1.1 Introdução

Pretendemos neste capítulo relembrar alguns conceitos básicos, que irão facilitar a compreensão dos métodos numéricos apresentados nos próximos capítulos. A maioria dos conceitos aqui apresentados são de álgebra linear e isso se deve ao fato de que os resultados da álgebra linear, em geral, e da teoria dos espaços vetoriais, em particular, na análise numérica é tão grande, que estudo pormenorizado desses assuntos cada vez mais se justifica. Assim maiores detalhes sobre os assuntos aqui abordados podem ser encontrados em livros de álgebra linear.

Para iniciar vamos examinar dois conjuntos que certamente já são conhecidos do leitor. O primeiro é o conjunto dos vetores da geometria, definidos através de segmentos orientados, e o outro é o conjunto das matrizes reais $m \times n$.

À primeira vista pode parecer que tais conjuntos não possuem nada em comum. Mas não é bem assim conforme mostraremos a seguir.

No conjunto dos vetores está definida uma adição dotada das propriedades comutativa, associativa, além da existência do elemento neutro (vetor nulo) e do oposto.

Além disso, podemos multiplicar um vetor por um número real. Essa multiplicação tem as seguintes propriedades (já certamente vista por você no seu curso):

$$\begin{aligned} \alpha(u+v) &= \alpha u + \alpha v \ , \\ (\alpha+\beta)u &= \alpha u + \beta u \ , \\ (\alpha\beta)u &= (\alpha\beta u) \ , \\ 1 \cdot u &= u \ , \end{aligned}$$

onde $u, \ v$ são vetores e $\alpha, \ \beta$ são escalares quaisquer.

No conjunto das matrizes também está definida uma adição dotada também das propriedades associativa, comutativa, admite elemento neutro, a matriz nula, e toda matriz tem uma oposta.

Como vemos o comportamento do conjunto dos vetores e o das matrizes quanto à adição é o mesmo. Mas não param por aí as coincidências.

Pode-se também multiplicar uma matriz por um número real. Essa multiplicação apresenta as mesmas propriedades que as destacadas para o caso de vetor, ou seja, valem as seguintes igualdades:

$$\begin{aligned} \alpha(A+B) &= \alpha A + \alpha B \ , \\ (\alpha+\beta)A &= \alpha A + \beta A \ , \\ (\alpha\beta)A &= (\alpha\beta A) \ , \\ 1 \cdot A &= A \ , \end{aligned}$$

onde $A,\ B$ são matrizes e $\alpha,\ \beta$ são escalares quaisquer.

Logo o conjunto dos vetores e o das matrizes apresentam uma certa coincidência estrutural no que se refere a um par importante de operações definidas sobre eles. Nada então mais lógico que estudar simultaneamente o conjunto dos vetores, das matrizes e todos os conjuntos que apresentem a mesma estrutura acima apontada.

1.2 Espaço Vetorial

Seja E um conjunto e seja K um corpo. Suponhamos que em E esteja definida uma operação de adição:

$$(x,y) \in E \times E \ \rightarrow x + y \in E \ ,$$

e que esteja definida uma operação entre os elementos de K e os elementos de E (chamada multiplicação por escalar):

$$(\alpha, x) \in K \times E \ \rightarrow \ \alpha x \in E \ .$$

Então E é um K**-espaço vetorial**, em relação a essas operações, se as seguintes condições estiverem satisfeitas:

$$\begin{array}{ll}
A_1) & (x+y)+z \ = \ x+(y+z), \ \forall x,y,z \in E \ , \\
A_2) & x+y \ = \ y+x, \ \forall x,y \in E \ , \\
A_3) & \exists \ 0(zero) \in E \ / \ x+0 \ = \ x, \ \forall x \in E \ , \\
A_4) & \forall x \in E, \ \exists -x \in E \ / \ x+(-x) \ = \ 0 \ , \\
M_1) & \alpha(x+y) \ = \ \alpha x + \alpha y, \ \forall \alpha \in K, \ \forall x,y \in E \ , \\
M_2) & (\alpha+\beta)x \ = \ \alpha x + \beta x, \ \forall \alpha,\beta \in K, \ \forall x,y \in E \ , \\
M_3) & (\alpha\beta)x \ = \ (\alpha\beta x), \ \forall \ \alpha,\beta \in K, \ \forall x \in E \ , \\
M_4) & 1 \cdot x \ = \ x, \ \forall \ x \in E \ .
\end{array}$$

O leitor deverá lembrar-se sempre de que, na definição acima, não se especifica nem a natureza dos vetores nem das operações. Assim qualquer conjunto que satisfaça as oito condições acima especificada será um espaço vetorial.

Definição 1.1 - *Seja E um K-espaço vetorial. Os vetores $v_1, v_2, \ldots, v_k \in E$ são* **linearmente dependentes** *sobre K, se existem escalares $\alpha_1, \alpha_2, \ldots, \alpha_k \in K$, nem todos nulos, tais que:*

$$\alpha_1 \ v_1 \ + \ \alpha_2 \ v_2 \ + \ldots + \ \alpha_k \ v_k \ = \ 0 \ .$$

Observamos que essa relação é sempre válida se os $\alpha_i, \ i = 1, 2, \ldots, k$ são todos iguais a zero. Nesse caso dizemos que os vetores são **linearmente independentes**.

Definição 1.2 - *Um K-espaço vetorial tem* **dimensão n** *se:*

a) existem **n** *vetores linearmente independentes;*

b) $(\mathbf{n}+\mathbf{1})$ *vetores são sempre linearmente dependentes.*

Definição 1.3 - *Qualquer conjunto de n vetores linearmente independentes é chamado* **base** *de um K-espaço vetorial de dimensão* **n***.*

Assim, qualquer vetor do espaço pode ser representado como combinação linear dos vetores da base.

Mudança de Base

Estudaremos inicialmente mudança de base em um espaço vetorial bi-dimensional, e a seguir, em um espaço de dimensão n.

a) Seja $E = I\!R^2$. Sejam $B_1 = \{e_1, e_2\}$ uma base de E e $v \in E$, como mostrados na Figura 1.1.

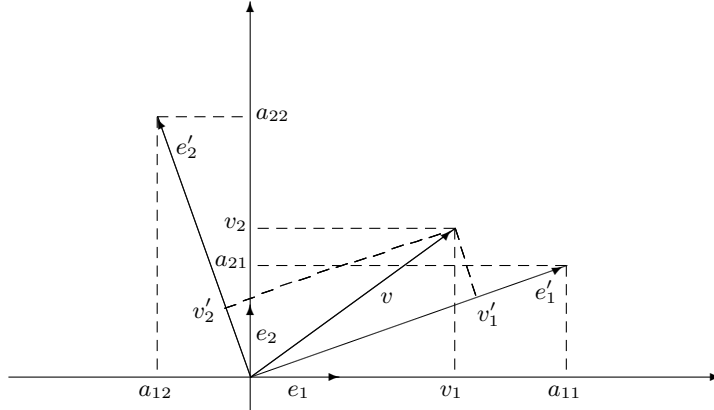



Figura 1.1

Então v se exprime de maneira única como combinação linear dos elementos de B_1, isto é, existem escalares v_1, v_2 (elementos de K) tais que:

$$v = v_1 \, e_1 + v_2 \, e_2 \, , \tag{1.1}$$

(onde os escalares v_1, v_2 são as coordenadas de v na base B_1).

Seja $B'_1 = \{e'_1, e'_2\}$, como mostrado na Figura 1.1, uma outra base de E. Analogamente, podemos escrever:

$$v = v'_1 \, e'_1 + v'_2 \, e'_2 \, . \tag{1.2}$$

Desejamos saber como, dadas as coordenadas de v na base B_1 (aqui denominada **base antiga**), poderemos determinar as coordenadas de v na base B'_1 (aqui denominada **base nova**). Sendo e'_1, e'_2 elementos de E podemos, em particular, escrever cada um deles como combinação linear dos elementos da base B_1. Assim:

$$\begin{array}{rcl} e'_1 & = & a_{11} \, e_1 + a_{21} \, e_2 \, , \\ e'_2 & = & a_{12} \, e_1 + a_{22} \, e_2 \, . \end{array} \tag{1.3}$$

isto é, cada vetor da base nova se exprime de maneira única como combinação linear dos vetores da base antiga.

Assim, em virtude de (1.1), (1.2) e (1.3) temos:

$$\begin{array}{rcl} v & = & v_1 \, e_1 + v_2 \, e_2 = v'_1 \, e'_1 + v'_2 \, e'_2 \\ & = & v'_1 \, (a_{11} \, e_1 + a_{21} \, e_2) + v'_2 \, (a_{12} \, e_1 + a_{22} \, e_2) \\ & = & (v'_1 \, a_{11} + v'_2 \, a_{12}) \, e_1 + (v'_1 \, a_{21} + v'_2 \, a_{22}) \, e_2 \, . \end{array}$$

Como as coordenadas de um vetor em relação a uma determinada base são únicas, podemos igualar os coeficientes. Assim, obtemos o sistema linear:

$$\left\{ \begin{array}{rcl} v_1 & = & v'_1 \, a_{11} + v'_2 \, a_{12} \\ v_2 & = & v'_1 \, a_{21} + v'_2 \, a_{22} \end{array} \right.$$

ou na forma matricial:

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} v_1' \\ v_2' \end{pmatrix} , \tag{1.4}$$

ou ainda:

$$v = A\, v' . \tag{1.5}$$

O sistema (1.4), possui sempre uma e uma só solução v_1', v_2', pelo fato de B_1 e B_1' serem bases de E.

Então, conhecidas, na base antiga, as coordenadas v_1, v_2 de v e as coordenadas de cada um dos vetores e_1', e_2', na base antiga, podemos determinar as coordenadas v_1', v_2' de v na base nova usando (1.4).

Sendo A não singular, $(det(A) \neq 0)$, existe a inversa A^{-1} de A. Assim, pré-multiplicando (1.5) por A^{-1}, obtemos:

$$v' = A^{-1}\, v . \tag{1.6}$$

A equação matricial (1.6) mostra como calcular as coordenadas de v na base antiga quando conhecidas as coordenadas de v na base nova.

Exemplo 1.1 - *Seja $v = (2,4)^t$ na base $\{(1,2)^t, (2,3)^t\}$. Calcular as coordenadas de v na base $\{(1,3)^t, (1,4)^t\}$.*

Solução: De (1.3), temos:

$$\begin{array}{rcl} (1,3)^t & = & a_{11}\,(1,2)^t + a_{21}\,(2,3)^t , \\ (1,4)^t & = & a_{12}\,(1,2)^t + a_{22}\,(2,3)^t . \end{array}$$

Da primeira equação, obtemos o sistema:

$$\left\{ \begin{array}{rcl} a_{11} + 2\,a_{21} & = & 1 \\ 2\,a_{11} + 3\,a_{21} & = & 3 \end{array} \right.$$

cuja solução é: $a_{11} = 3,\ a_{21} = -1$. De maneira análoga, da segunda equação, obtemos:

$$\left\{ \begin{array}{rcl} a_{12} + 2\,a_{22} & = & 1 \\ 2\,a_{12} + 3\,a_{22} & = & 4 \end{array} \right.$$

cuja solução é: $a_{12} = 5,\ a_{22} = -2$. Substituindo os valores conhecidos em (1.4), segue que:

$$\begin{pmatrix} 2 \\ 4 \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ -1 & -2 \end{pmatrix} \begin{pmatrix} v_1' \\ v_2' \end{pmatrix} .$$

cuja solução é: $v_1' = 24,\ v_2' = -14$. Assim, $v = (24, -14)^t$ na base $\{(1,3)^t, (1,4)^t\}$.

Veremos agora, mudança de base em um K-espaço vetorial E de dimensão **n**.

b) Seja $E = I\!R^n$. Sejam $\{e_1, e_2, \ldots, e_n\}$, $\{e_1', e_2', \ldots, e_n'\}$ bases de E e $v \in E$. Então, podemos escrever:

$$v = \sum_{i=1}^{n} v_i e_i = \sum_{j=1}^{n} v_j' e_j' .$$

Mas $e_1', e_2', \ldots, e_n'$ são elementos de E, e portanto podem ser expressos em relação a base $\{e_1, e_2, \ldots, e_n\}$. Logo:

$$e_j' = \sum_{i=1}^{n} a_{ij} e_i \quad , \quad j = 1, 2, \ldots, n .$$

Então temos:

$$\begin{aligned} v &= \sum_{i=1}^{n} v_i \; e_i \; = \; \sum_{j=1}^{n} v'_j \; e'_j \\ &= \sum_{j=1}^{n} v'_j \left(\sum_{i=1}^{n} a_{ij} \; e_i \right) = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} \; v'_j \right) \; e_i \; , \; \Rightarrow \; v_i \; = \; \sum_{j=1}^{n} a_{ij} v'_j \; . \end{aligned}$$

Assim, na forma matricial, podemos escrever:

$$\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} v'_1 \\ v'_2 \\ \vdots \\ v'_n \end{pmatrix} .$$

ou

$$v \; = \; A \, v' \qquad \text{e} \qquad v' \; = \; A^{-1} \; v \; .$$

Exercícios

1.1 - *Seja* $v = (2, \; 3, \; 4)^t$ *na* **base canônica**, *isto é, na base:*
$\{(1, \; 0, \; 0)^t \; , \; (0, \; 1, \; 0)^t \; , \; (0, \; 0, \; 1)^t\}$.
Calcular as coordenadas de v *na base:*
$\{(1, \; 1, \; 1)^t \; , \; (1, \; 1, \; 0)^t \; , \; (1, \; 0, \; 0)^t\}$.

1.2 - *Seja* $v = 3 \; b_1 + 4 \; b_2 + 2 \; b_3$*, onde:*
$b_1 = (1, \; 1, \; 0)^t \; , \; b_2 = (-1, \; 1, \; 0)^t \; , \; b_3 = (0, \; 1, \; 1)^t$.
Calcular as coordenadas de v *na base:*
$f_1 = (1, \; 1, \; 1)^t \; , \; f_2 = (1, \; 1, \; 0)^t \; , \; f_3 = (1, \; 0, \; 0)^t$.

1.3 - *Seja* $K_n(x) = \{P_r(x) \; / \; r \leq n\}$ *o espaço vetorial de todos os polinômios de grau* $\leq n$*. A* **base canônica** *para o espaço dos polinômios é* $\{1, \; x, \; x^2, \; \ldots\}$*. Seja* $P_3 = 3 + 4 \; x^2 + 2 \; x^3$ *e* $B_1 = \{5, \; x - 1, \; x^2 - 5 \; x + 3, \; x^3 - 4\}$ *uma outra base. Calcular as coordenadas de* P_3 *em relação à base* B_1*.*

1.4 - *Sejam* $B_1 = \{5, \; x - 1, \; x^2 - 3 \; x\}$ *e* $B_2 = \{8, \; 3 \; x + 2, 5 \; x^2 - 3 \; x\}$ *bases de* $K_2(x)$*. Seja* $P_2(x) = 8\{5\} + 4\{x - 1\} + 3\{x^2 - 3x\}$*. Calcular as coordenadas de* $P_2(x)$ *em relação à base* B_2*.*

1.5 - *Dado o polinômio* $P_3(x) = 20 \; x^3 + 8 \; x^2 - 14 \; x + 28$ *exprimí-lo como combinação linear dos polinômios da sequência:*

$$\begin{aligned} &Q_3(x) = 5 \; x^3 - 7 \; x + 12, \\ &Q_2(x) = -4 \; x^2 + 8 \; x, \\ &Q_1(x) = 6 \; x - 1, \\ &Q_0(x) = 5. \end{aligned}$$

Espaço Vetorial Euclidiano

Vamos definir aqui importantes noções de produto escalar e de ortogonalidade, visando introduzir, entre outras coisas o conceito de comprimento e distância.

Produto Escalar

Seja E um espaço vetorial real. Sejam x, y elementos de E.

Definição 1.4 - *Chama-se* **produto escalar** *(ou* **produto interno***) de x por y, em símbolo, (x, y), qualquer função definida em $E \times E$ com valores em $I\!R$ satisfazendo as seguintes propriedades:*

$$\begin{array}{ll} P_1) & (x,y) \ = \ (y,x), \quad \forall x,y \in E \ , \\ P_2) & (x+y,z) \ = \ (x,z) \ + \ (y,z), \quad \forall x,y,z \in E \ , \\ P_3) & (\lambda x,y) \ = \ \lambda(x,y), \quad \forall \lambda \in I\!R, \quad \forall x,y \in E \ , \\ P_4) & (x,x) \ \geq \ 0 \text{ e } (x,x) \ = \ 0 \text{ se e somente se } x = \theta(nulo). \end{array}$$

Um espaço vetorial real E, onde está definido um produto escalar é chamado **espaço euclidiano real**.

Daremos a seguir alguns exemplos de produto escalar.

Exemplo 1.2 - *Seja $E = I\!R^2$. Sejam $x = (x_1, x_2)^t$; $y = (y_1, y_2)^t$. Mostrar que, definindo:*

$$(x,y) \ = \ x_1 \ y_1 \ + \ x_2 \ y_2 \ . \tag{1.7}$$

o $I\!R^2$ torna-se um espaço euclidiano real.

Solução: Devemos mostrar que as condições P_1, P_2, P_3 e P_4 estão satisfeitas, isto é, que (1.7) é um produto escalar bem definido no $I\!R^2$. De fato:

$$\begin{array}{llll} P_1) & (x,y) & = & x_1y_1 + x_2y_2 = y_1x_1 + y_2x_2 = (y,x). \\ P_2) & (x+y,z) & = & (x_1+y_1)z_1 + (x_2+y_2)z_2 = x_1z_1 + y_1z_1 + x_2z_2 + y_2z_2 \\ & & = & (x_1z_1 + x_2z_2) + (y_1z_1 + y_2z_2) = (x,z) + (y,z). \\ P_3) & (\lambda \ x,y) & = & \lambda x_1y_1 + \lambda x_2y_2 = \lambda(x_1y_1 + x_2y_2) = \lambda(x,y). \\ P_4) & (x,x) & = & x_1^2 + x_2^2 \ \geq \ 0 \quad (evidente). \\ & (x,x) & = & x_1^2 + x_2^2 = 0 \Leftrightarrow x_i^2 = 0 \Leftrightarrow \ x_i \ = \ 0, \forall_i \ \Leftrightarrow x = \theta. \end{array}$$

Logo, (1.7) é uma boa definição de produto escalar.

Nos próximos exemplos, a verificação de que as condições P_1, P_2, P_3 e P_4 são satisfeitas, fica como exercício.

Exemplo 1.3 - *Seja $E = I\!R^n$. Para $x, y \in E$, isto é, $x = (x_1, \ x_2, \ \ldots, \ x_n)^t$, e $y = (y_1, \ y_2, \ \ldots, \ y_n)^t$, definimos:*

$$(x,y) \ = \ \sum_{i=1}^{n} \ x_i \ y_i \ , \tag{1.8}$$

como um produto escalar no $I\!R^n$. (1.8) é chamado de produto escalar usual no $I\!R^n$. Também,

$$(x,y) \ = \ \sum_{i=1}^{n} \ w_i \ x_i \ y_i, \tag{1.9}$$

com w_i fixados e positivos, define no $I\!R^n$ um produto escalar.

Assim, tanto (1.8) como (1.9) transformam o $I\!R^n$ num espaço euclidiano real.

Exemplo 1.4 - *Seja $E = C[a, b]$ o espaço vetorial das funções contínuas reais definidas sobre o intervalo limitado fechado $[a, b]$. Se para $f, g \in C[a, b]$ definimos:*

$$(f,g) \ = \ \int_a^b f(x) \ g(x) dx, \tag{1.10}$$

tal espaço torna-se um espaço euclidiano real. (1.10) é chamado de produto escalar usual em $C[a, b]$.

Em particular, se $f(x) = P_k(x)$ e $g(x) = P_j(x)$, com $k,\ j\ \leq\ n$, são polinômios de grau $\leq n$, a equação (1.10) define um produto escalar em $K_n = \{P_r(x)\ /\ r \leq n\}$, (espaço vetorial dos polinômios de grau $\leq n$).

Exemplo 1.5 - *Seja* $E = K_n(x) = \{P_r(x)\ /\ r \leq n\}$. *Sejam* $a \leq x_0 < x_1 < \ldots < x_m \leq b,\ m+1$ *pontos distintos, com* $m \geq n$. *Definimos:*

$$(P_i(x), P_j(x)) \ = \ \sum_{k=0}^{m} P_i\,(x_k)\,P_j\,(x_k)\,. \tag{1.11}$$

como um produto escalar K_n.

Esse último exemplo mostra uma outra maneira de se transformar $K_n(x)$ num espaço euclidiano real, maneira esta que será útil em problemas de aproximação de funções pelo método dos mínimos quadrados, no caso discreto.

Ortogonalidade

Seja E um espaço euclidiano real. Sejam x, y elementos de E.

Definição 1.5 - *Dizemos que* x *é* **ortogonal** *a* y, *em símbolo,* $x \perp y$, *se e somente se* $(x, y) = 0$.

Observe que $(x, \theta) = (\theta, x) = 0$ qualquer que seja x, onde θ é o vetor nulo.

Exemplo 1.6 - *No espaço* $E = C[-\pi, \pi]$, *com* $(f, g) = \int_{-\pi}^{\pi} f(x)\ g(x)\ dx$, *verificar se* $sen\ x$ *e* $\cos\ x$ *são ortogonais.*

Solução: Temos:

$$(sen\ x, cos\ x) \ = \ \int_{-\pi}^{\pi} sen\ x\ cos\ x\ dx \ = \ \left.\frac{sen^2\ x}{2}\right]_{-\pi}^{\pi} \ = \ 0\ .$$

Assim, $sen\ x$ e $cos\ x$ são ortogonais em E.

Exemplo 1.7 - *Em* $E = I\!R^3$, *com o produto escalar usual, verificar se os vetores:* $f_1 = \left(\frac{1}{\sqrt{3}},\ \frac{1}{\sqrt{3}},\ \frac{1}{\sqrt{3}}\right)^t$ *e* $f_2 = \left(\frac{1}{\sqrt{2}},\ -\frac{1}{\sqrt{2}},\ 0\right)^t$ *são ortogonais.*

Solução: Temos:

$$\begin{aligned} (f_1, f_2) \ &= \ \frac{1}{\sqrt{3}} \times \frac{1}{\sqrt{2}} \ + \ \frac{1}{\sqrt{3}} \ \times \ \left(-\frac{1}{\sqrt{2}}\right) \ + \ \frac{1}{\sqrt{3}} \ \times \ 0 \\ &= \ \frac{1}{\sqrt{6}} \ - \ \frac{1}{\sqrt{6}} \ + \ 0 \ = \ 0. \end{aligned}$$

Logo, f_1 e f_2 são ortogonais em E.

Teorema 1.1 - *Os vetores* $v_1, v_2, \ldots, v_m$ *tais que:*

a) $v_i \ \neq \ \theta,\ i \ = \ 1, 2, \ldots, m$;
b) $(v_i, v_j) \ = \ 0$, para $i \ \neq \ j$;

são sempre linearmente independentes.

Dito de outro modo:*os vetores não nulos $v_1, v_2, \ldots, v_m$, dois a dois ortogonais, são sempre linearmente independentes.*

Prova: Devemos provar que:

$$\alpha_1 v_1 \ + \ \alpha_2 v_2 \ + \ldots + \ \alpha_m v_m \ = \ 0 \tag{1.12}$$

$$\Rightarrow \ \alpha_1 \ = \ \alpha_2 \ = \ \ldots \ = \ \alpha_m \ = \ 0.$$

Em virtude de (1.12) podemos escrever, sucessivamente, para cada $i \ = \ 1, 2, \ldots, m$:

$$(v_i \ , \ \alpha_1 v_1 \ + \ \alpha_2 v_2 \ + \ldots + \ \alpha_i v_i \ + \ldots + \ \alpha_m v_m) \ = \ (v_i, 0) \ = \ 0,$$

ou seja:

$$\alpha_1 \, (v_i, v_1) \ + \ \alpha_2 \, (v_i v_2) \ + \ldots + \ \alpha_i \, (v_i, v_i) \ + \ldots + \ \alpha_m \, (v_i, v_m) \ = \ 0.$$

onde aplicamos P_2 e P_3. Mas $(v_i, v_j) \ = \ 0 \ \ , \ \ i \ \neq \ j$. Daí, a igualdade acima se reduz a:

$$\alpha_i \, (v_i, v_i) \ = \ 0.$$

Mas sendo $v_i \ \neq \ \theta$, temos, usando P_4, que $(v_i, v_i) \neq 0$, para $i = 1, 2, \ldots, m$. Portanto, da última igualdade concluímos que,

$$\alpha_i \ = \ 0, \ i \ = \ 1, 2, \ldots, m.$$

Logo, os vetores $v_1, v_2, \ldots, v_m$ são linearmente independentes.

Definição 1.6 - *Seja E um espaço euclidiano de dimensão n. Se $f_1, f_2, \ldots, f_n$ são dois a dois ortogonais, ou seja, se $(f_i, f_j) = 0, \ i \neq j$, eles constituem uma base de E, que será chamada de* **base ortogonal**.

Teorema 1.2 - *A condição necessária e suficiente para que um vetor $v \in E$ seja ortogonal a um sub-espaço $E' \subset E$ é que v seja ortogonal a cada vetor $e_1, e_2, \ldots, e_n$ de uma base de E'.*

Prova: A condição é evidentemente necessária. Provemos a suficiência. Seja x um vetor qualquer de E'. Temos então:

$$x \ = \ \alpha_1 \ e_1 \ + \ \alpha_2 \ e_2 \ + \ \ldots \ + \ \alpha_n \ e_n,$$

desde que $e_1, e_2, \ldots, e_n$ é uma base de E'. Devemos mostrar que $v \perp x$. Assim:

$$\begin{aligned} (v, x) \ &= \ (v, \ \alpha_1 \ e_1 \ + \ \alpha_2 \ e_2 \ + \ldots + \ \alpha_n \ e_n) \\ &= \ \alpha_1 \, (v, e_1) \ + \ \alpha_2 \, (v, e_2) \ + \ldots + \ \alpha_n \, (v, e_n) \ = \ 0, \end{aligned}$$

desde que por hipótese, $v \perp \{e_1, e_2, \ldots, e_n\}$. Logo v é ortogonal a E'.

Teorema 1.3 - *Num espaço euclidiano real E quaisquer que sejam $x, y \in E$, temos:*

$$(x, y)^2 \ \leq \ (x, x) \ (y, y), \tag{1.13}$$

com igualdade válida se e somente se x e y são linearmente dependentes.

A desigualdade (1.13) é chamada **desigualdade de Schwarz**.

Prova: Tomemos o vetor $v = x + \lambda \ y$, onde λ é um número real qualquer. De P_4, resulta:

$$(x + \lambda \ y, x + \lambda \ y) \geq 0 \ ,$$

e usando P_2 e P_3, obtemos:

$$\lambda^2 (y, y) \ + \ 2\lambda (x, y) \ + \ (x, x) \ \geq \ 0 \ .$$

Para que o trinômio seja sempre ≥ 0 é necessário que $\Delta \leq 0$. Assim:

$$\begin{aligned} \Delta &= 4(x,y)^2 - 4(x,x)(y,y) \leq 0, \\ \Rightarrow & (x,y)^2 \leq (x,x)(y,y). \end{aligned}$$

Mostremos agora que a igualdade é válida se e somente se x e y são linearmente dependentes. Seja $x = \lambda\, y$. Então:

$$\begin{aligned} (x,y)^2 &= (\lambda y, y)^2 = [\lambda(y,y)]^2 = \lambda^2(y,y)^2 \\ &= \lambda^2(y,y)(y,y) = (\lambda y, \lambda y)(y,y) = (x,x)(y,y). \end{aligned}$$

Isto é, x e y linearmente dependentes $\Longrightarrow (x,y)^2 = (x,x)(y,y)$.

Suponhamos, agora que a igualdade seja válida em (1.13). O caso $y = \theta$ é trivial. Suponhamos $y \neq \theta$. Temos que $(x,y)^2 = (x,x)(y,y)$ é equivalente a:

$$(x + \lambda\, y, x + \lambda\, y) = 0 \text{ com } \lambda = -\frac{(x,y)}{(y,y)}.$$

Assim, de P_4, concluímos que $x + \lambda\, y = 0$. Ou seja $x = \dfrac{(x,y)}{(y,y)}\, y$, e isto quer dizer que x e y são linearmente dependentes.

Exercícios

1.6 - *Em relação ao produto escalar usual do $\mathbb{R}^3$, calcule (x,y) nos seguintes casos:*

a) $x = (1/2,\ 2,\ 1)^t$, $y = (4,\ 1,\ -3)^t$;

b) $x = (2,\ 1,\ 0)^t$, $y = (4,\ 0,\ 2)^t$;

1.7 - *Determinar $(f,g) = \int_0^1 f(t)g(t)dt$ para cada um dos seguintes pares de vetores de $K_2(t)$.*

a) $f(t) = t$, $g(t) = 1 - t^2$;

b) $f(t) = t - \frac{1}{2}$, $g(t) = \frac{1}{2} - \left(t - \frac{1}{2}\right)$;

1.8 - *Sejam $x = (x_1, x_2)^t$ e $y = (y_1, y_2)^t$ dois vetores quaisquer do $\mathbb{R}^2$. Mostre que:*

$$(x,y) = \frac{x_1 x_2}{a^2} + \frac{y_1 y_2}{b^2} ,$$

com $a, b \in \mathbb{R}$ fixos e não nulos define um produto escalar sobre o $\mathbb{R}^2$.

1.9 - *Considere no espaço vetorial $\mathbb{R}^2$ o produto escalar dado por: $(x,y) = x_1 y_1 + 2x_2 y_2$, para todo par de vetores $x = (x_1, x_2)^t$ e $y = (y_1, y_2)^t$. Verificar se x e y são ortogonais em relação a esse produto escalar nos seguintes casos:*

a) $x = (1,\ 1)^t$ e $y = (2,\ -1)^t$;

b) $x = (2,\ 1)^t$ e $y = (-1,\ 1)^t$;

b) $x = (3,\ 2)^t$ e $y = (2,\ -1)^t$;

1.10 - *Determine* m *de modo que sejam ortogonais os vetores* $x = (m+1,\ 2)^t$ *e* $y = (-1,\ 4)^t$ *em relação ao produto escalar usual do* $I\!R^2$.

1.11 - *Determinar* $f(x) \in K_2(x)$ *que seja ortogonal a* $g(x) = 1$ *e* $h(x) = t$, *em relação ao produto escalar dado por:*

$$(f,g) = \int_{-1}^{1} f(x)\ g(x)\ dx\ .$$

1.12 - *Considere no* $I\!R^3$ *o produto escalar usual. Determine* $m \in I\!R$ *de tal modo que os vetores* $u = (1,\ m+1,\ m)^t$, $v = (m-1,\ m,\ m+1)^t$, *sejam ortogonais.*

1.13 - *Sejam* $f(x) = x$, $g(x) = mx^2 - 1$ *e considere o produto escalar usual em* $C[0,1]$. *Determine o valor de* m, *para que* $f(x)$ *e* $g(x)$ *sejam ortogonais.*

Espaço Vetorial Normado

Vamos definir agora importantes definições de norma de vetor e de matriz. Com isso estaremos aptos a definir, quando oportuno, as noções de limite de uma sequência de vetores ou de matrizes, de grande utilidade, entre outros, no estudo de convergência de métodos iterativos de solução de sistemas lineares e do problema de erros de arredondamento nos processos de cálculo onde intervêm matrizes ou vetores.

Norma de Vetor

Definição 1.7 - *Chama-se* **norma** *de um vetor* x, *em símbolo,* $\| x \|$, *qualquer função definida num espaço vetorial* E, *com valores em* $I\!R$, *satisfazendo as seguintes condições:*

$N_1)$ $\| x \| \geq 0$ e $\| x \| = 0$ se e somente se $x = \theta$,
$N_2)$ $\| \lambda\, x \| = |\lambda|\ \| x \|$ para todo escalar λ
$N_3)$ $\| x + y \| \leq \| x \| + \| y \|$ (desigualdade triangular).

Um espaço vetorial E, onde está definida uma norma é chamado **espaço vetorial normado**.

Daremos a seguir alguns exemplos de norma no $I\!R^n$.

Exemplo 1.8 - *Seja* $E = I\!R^n$, *e seja* $x = (x_1, x_2, \ldots, x_n)^t$. *Mostrar que, definindo:*

$$\| x \|_E = \sqrt{\sum_{i=1}^{n} x_i^2}, \tag{1.14}$$

o $I\!R^n$ *torna-se um espaço vetorial normado.*

Solução: Vamos mostrar que as condições N_1, N_2 e N_3 estão satisfeitas, isto é, que (1.14) é uma norma

bem definida no $I\!R^n$. De fato:

$$N_1) \quad \| x \|_E = \sqrt{\sum_{i=1}^{n} x_i^2} \geq 0 \quad (evidente).$$

$$\| x \|_E = \sqrt{\sum_{i=1}^{n} x_i^2} = 0 \Leftrightarrow \sum_{i=1}^{n} x_i^2 = 0 \Leftrightarrow x_i = 0, \forall_i \Leftrightarrow x = \theta.$$

$$N_2) \quad \| \lambda x \|_E = \sqrt{\sum_{i=1}^{n} \lambda^2 x_i^2} = \sqrt{\lambda^2 \sum_{i=1}^{n} x_i^2} = |\lambda| \sqrt{\sum_{i=1}^{n} x_i^2} = |\lambda| \, \| x \|_E \, .$$

$$\begin{aligned} N_3) \quad \| x + y \|_E^2 &= \sum_{i=1}^{n} (x_i + y_i)^2 = (x_1 + y_1)^2 + (x_2 + y_2)^2 + \ldots + (x_n + y_n)^2 \\ &= x_1^2 + 2x_1 y_1 + y_1^2 + x_2^2 + 2x_2 y_2 + y_2^2 + \ldots + x_n^2 + 2x_n y_n + y_n^2 \\ &= \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2 \\ &\leq \sum_{i=1}^{n} x_i^2 + 2 \sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2} + \sum_{i=1}^{n} y_i^2 \, , \end{aligned}$$

onde usamos a desigualdade de Schwarz, isto é:

$$\sum_{i=1}^{n} x_i y_i \leq \sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}.$$

Portanto,

$$\begin{aligned} \| x + y \|_E^2 &\leq \| x \|_E^2 + 2 \| x \|_E \| y \|_E + \| y \|_E^2 \\ &= (\| x \|_E + \| y \|_E)^2 \, . \end{aligned}$$

Assim: $\| x + y \|_E^2 \leq (\| x \|_E + \| y \|_E)^2$. Extraindo-se a raiz quadrada de ambos os membros, temos que: $\| x + y \|_E \leq \| x \|_E + \| y \|_E$.

Logo, (1.14) é uma boa definição de norma.

No próximo exemplo, a verificação de que as condições N_1, N_2 e N_3 são satisfeitas, fica como exercício.

Exemplo 1.9 - *Seja $E = I\!R^n$, e seja $x = (x_1, x_2, \ldots x_n)^t$. Definimos então:*

$$\| x \|_\infty = \max_{1 \leq i \leq n} |x_i| \, ,$$

$$\| x \|_1 = \sum_{i=1}^{n} |x_i| \, ,$$

$$\| x \| = \sqrt{(x, x)} \, ,$$

como normas no $I\!R^n$.

Observações:

1) $\| x \| = \sqrt{(x, x)}$ corresponde à noção intuitiva de **comprimento ou módulo de um vetor**.

2) Se usarmos a definição usual de produto escalar no $I\!R^n$, isto é, se usarmos (1.8), então: $\| x \| = \sqrt{(x,x)} = \sqrt{\sum_{i=1}^{n} x_i^2} = \| x \|_E$.

Exemplo 1.10 - *Seja* $x = (-1,\ 10,\ 3,\ 4,\ -20)^t$. *Calcular* $\| x \|_E$, $\| x \|_\infty$ *e* $\| x \|_1$.

Solução: Aplicando a definição de cada uma das normas, obtemos:

$$\begin{array}{lcl} \| x \|_E & = & \sqrt{(-1)^2 + (10)^2 + 3^2 + 4^2 + (-20)^2} \simeq 22.93, \\ \| x \|_\infty & = & max\ (|-1|, |10|, |3|, |4|, |-20|) = 20, \\ \| x \|_1 & = & |-1| + |10| + |3| + |4| + |-20| = 38. \end{array}$$

Como você pode observar a aplicação de cada uma das normas definidas anteriormente fornece um resultado diferente. Entretanto, no $I\!R^n$, todas as normas são equivalentes.

Definição 1.8 - *Duas normas* $\| \cdot \|_a$ *e* $\| \cdot \|_b$ *são* **equivalentes** *se existem constantes* k_1 *e* k_2 *tais que:*

$$k_1 \ \| x \|_a \leq \| x \|_b \leq \ k_2 \ \| x \|_a \quad , \quad \forall\ x \in E. \tag{1.15}$$

Exemplo 1.11 - *Como exemplos de normas equivalentes, no* $I\!R^n$, *temos:*

$$\begin{array}{ll} a) & \| x \|_\infty \leq \| x \|_1 \leq n \ \| x \|_\infty , \\ b) & \| x \|_\infty \leq \| x \|_E \leq \sqrt{n} \ \| x \|_\infty , \\ c) & \frac{1}{n} \| x \|_1 \leq \| x \|_E \leq \sqrt{x} \ \| x \|_1 . \end{array}$$

Vamos verificar que o item a) é verdadeiro; a verificação das demais fica como exercício.

Solução: Temos:

$$\begin{array}{lcl} \| x \|_\infty & = & \max\limits_{1 \leq i \leq n} |x_i| = \max\{|x_1|, |x_2|, \ldots, |x_n|\} \\ & = & |x_k| \leq |x_k| + \sum\limits_{i=1}^{k-1} |x_i| + \sum\limits_{i=k+1}^{n} |x_i| = \sum\limits_{i=1}^{n} |x_i| = \| x \|_1 \\ & = & |x_1| + |x_2| + \ldots + |x_n| \leq \{\underbrace{|x_k| + |x_k| + \ldots + |x_k|}_{n \text{ vezes}}\} \\ & = & n|x_k| = n \max\limits_{1 \leq i \leq n} |x_i| = n \ \| x \|_\infty . \end{array}$$

Teorema 1.4 - *A desigualdade de Schwarz* (1.13) *pode ser escrita como:*

$$|(x,y)| \leq \| x \| \ \| y \| \ . \tag{1.16}$$

Prova: A prova deste teorema fica como exercício.

Um vetor x, de E, é **unitário** se seu comprimento é igual a 1, isto é, se $\| x \| = 1$.

Definição 1.9 - *Seja* E *um espaço euclidiano de dimensão* n. *Os vetores* $f_1, f_2, \ldots, f_n$ *formam uma* **base ortonormal** *de* E *se eles forem vetores ortonormais, ou seja, se:*

$$(f_i, f_j) = \delta_{ij} = \left\{ \begin{array}{ll} 1 & \text{se} \quad i = j, \\ 0 & \text{se} \quad i \neq j. \end{array} \right.$$

Assim uma sequência de vetores é ortonormal se cada um dos seus elementos tem norma 1 e dois quaisquer distintos dentre eles são ortogonais.

Teorema 1.5 - *Num espaço euclidiano, um conjunto ortornormal de vetores é sempre linearmente independente.*

Prova: (análoga ao do Teorema 1.1)).

Definição 1.10 - *Seja E um espaço euclidiano. Dados os vetores x e $y \in E$, definimos* **distância** *entre x e y, o comprimento do vetor $x - y$, isto é:*

$$d(x,y) \ = \| \ x - y \ \| \ \rightarrow \ d(x,y) \ = \ \sqrt{(x-y, x-y)}.$$

Temos assim uma aplicação $d : E \times E \ \rightarrow \ I\!\!R$, que satisfaz as seguintes condições:

$$\begin{array}{ll} D_1) & d(x,y) \geq 0 \text{ e } d(x,y) = 0 \text{ se e somente se } x = y \ , \\ D_2) & d(x,y) \ = \ d(y,x) \ , \ \forall x,y \in E \ , \\ D_3) & d(x,y) \leq d(x,z) + d(z,y) \ , \ \forall x,y,z \ \in E \ . \end{array}$$

Norma de Matriz

Como já dissemos anteriormente, o conjunto das matrizes $(n \times n)$, com as operações de soma de matrizes e produto de um escalar por uma matriz forma um espaço vetorial E de dimensão n^2. Podemos então falar em norma de uma matriz $A \in E$. Observe então que no caso de matrizes, vale a mesma definição de norma de vetor , isto é:

Definição 1.11 - *Chama-se* **norma** *de uma matriz A, em símbolo, $\| \ A \ \|$, qualquer função definida no espaço vetorial das matrizes $n \times n$, com valores em $I\!\!R$, satisfazendo as seguintes condições:*

$$\begin{array}{ll} M_1) & \| \ A \ \| \ \geq \ 0 \text{ e } \ \| \ A \ \| \ = \ 0 \text{ se e somente se } A \ = \ \theta (\text{matriz nula}) \ , \\ M_2) & \| \ \lambda \ A \ \| \ = \ |\lambda| \ \ \| \ A \ \| \ \text{ para todo escalar } \lambda \ , \\ M_3) & \| \ A + B \ \| \ \leq \ \| \ \ A \ \| \ \ + \| \ B \ \| \ \ (\text{desigualdade triangular}) \ . \end{array}$$

Daremos a seguir alguns exemplos de norma de matrizes. A verificação de que são normas bem definidas no espaço vetorial das matrizes $n \times n$, fica a cargo do leitor.

Exemplo 1.12 - *Seja A uma matriz $(n \times n)$. Definimos então:*

$$a) \quad \| \ A \ \|_\infty \ = \ \max_{1 \leq i \leq n} \ \sum_{j=1}^{n} \ |a_{ij}| \quad (\textit{norma linha}) \ ;$$

$$b) \quad \| \ A \ \|_1 \ = \ \max_{1 \leq j \leq n} \ \sum_{i=1}^{n} \ |a_{ij}| \quad (\textit{norma coluna}) \ ;$$

$$c) \quad \| \ A \ \|_E \ = \ \sqrt{\sum_{i,j=1}^{n} a_{ij}^2} \quad (\textit{norma euclidiana}) \ .$$

Para essas normas vale: $\| \ AB \ \| \leq \| \ A \ \| \| \ B \ \|$. (Prove).

Exemplo 1.13 - *Seja*

$$A = \begin{pmatrix} 3 & 2 & -1 \\ 6 & 3 & 4 \\ -1 & 2 & 1 \end{pmatrix} .$$

Calcular $||A||_\infty$, $||A||_1$, $||A||_E$.

Solução: Usando cada uma das definições dadas anteriormente, obtemos:

$$\begin{aligned} ||A||_\infty &= |6| + |3| + |4| = 13 \ , \\ ||A||_1 &= |3| + |6| + |-1| = 10 \ , \\ ||A||_E &= (9+4+1+36+9+16+1+4+1)^{1/2} = 9 \ . \end{aligned}$$

Como no caso de vetor, as normas de matrizes também são equivalentes, isto é, satisfazem uma relação do tipo (1.15), com o vetor x substituído pela matriz A. A verificação das desigualdades no próximo exemplo fica como exercício.

Exemplo 1.14 - *Como exemplos de normas equivalentes, no espaço vetorial das matrizes de ordem n, temos:*

$$\begin{aligned} &a) \quad \frac{1}{n} \parallel A \parallel_\infty \leq \parallel A \parallel_E \leq \sqrt{n} \parallel A \parallel_\infty \ , \\ &b) \quad \frac{1}{n} \parallel A \parallel_1 \leq \parallel x \parallel_E \leq \sqrt{n} \parallel x \parallel_1 \ , \\ &c) \quad \parallel A \parallel_\infty \leq n \parallel A \parallel_1 \ , \\ &d) \quad \parallel A \parallel_1 \leq n \parallel A \parallel_\infty \ . \end{aligned}$$

Definição 1.12 - *Dada uma norma de vetor, podemos definir uma norma de matriz, que será chamada de* **subordinada** *a ela do seguinte modo:*

$$\parallel A \parallel = \sup_{\parallel x \parallel = 1} \parallel Ax \parallel \ .$$

Observe que a norma de matriz assim definida pode ser interpretada como sendo o comprimento do maior vetor no conjunto imagem $\{Ax\}$ da esfera unitária $\{x \ / \ \parallel x \parallel = 1\}$ pela transformação $x \to Ax$.

Definição 1.13 - *Se uma norma de matriz e uma norma de vetor estão relacionadas de tal modo que a desigualdade:*

$$\parallel Ax \parallel \leq \parallel A \parallel \parallel x \parallel \ ,$$

é satisfeita para qualquer x, então dizemos que as duas normas são **consistentes**.

Note que existe um vetor x_0 tal que: $\parallel Ax \parallel = \parallel A \parallel \parallel x \parallel$. Nestas condições: $\parallel A \parallel = mink$ tal que $\parallel Ax \parallel \leq k \ \parallel x \parallel$.

Exercícios

1.14 - *Considere os vetores do $I\!R^6$: $x = (1,\ 2,\ 0,\ -1,\ 2,\ -10)^t$ e $y = (3,\ 1,\ -4,\ 12,\ 3,\ 1)^t$. Calcule a norma de cada um desses vetores usando as normas definidas no exemplo 1.9.*

1.15 - *No espaço vetorial $I\!R^4$, munido do produto escalar usual, sejam $x = (1,\ 2,\ 0,\ 1)^t$ e $y = (3,\ 1,\ 4,\ 2)^t$. Determine: $(x,y), \parallel x \parallel, \parallel y \parallel, d(x,y)$ e $\dfrac{x+y}{\parallel x+y \parallel}$.*

1.16 - *Prove que num espaço euclidiano normado:*

a) $\parallel x+y \parallel^2 + \parallel x-y \parallel^2 = 2(\parallel x \parallel^2 \parallel +y \parallel^2)$,

b) $| \parallel x \parallel - \parallel y \parallel | \leq \parallel x-y \parallel$.

1.17 - *Sejam u e v vetores de um espaço euclidiando tais que $\| u \| = 1, \| v \| = 1$ e $\| u - v \| = -2$. Determine (u, v).*

1.18 - *Considere as seguintes matrizes:*

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}; \quad B = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & 3 & 2 \end{pmatrix}; \quad C = \begin{pmatrix} 2 & 1 & 3 & -1 \\ 4 & 3 & 8 & 2 \\ 6 & 7 & 10 & 1 \\ 3 & -1 & 0 & 1 \end{pmatrix}.$$

Calcule a norma de cada uma delas usando as normas definidas no exemplo 1.12.

1.3 Processo de Gram-Schmidt

Em diversos problemas relacionados com espaço vetorial, a escolha de uma base para o espaço fica a critério da pessoa que se propôs a resolver o problema. É claro que sempre a melhor estratégia será escolher a base que melhor simplifique os cálculos. Em espaços euclidianos, tem-se muitas vezes o caso em que a melhor escolha da base é aquela onde todos os seus vetores são mutuamente ortogonais ou ortonormais.

Vimos anteriormente que uma sequência ortonormal de vetores é sempre linearmente independente. Vamos agora mostrar que é sempre possível construir, a partir de uma sequência de vetores linearmente independentes $\{f_1, f_2, \ldots, f_n\}$, uma sequência ortogonal $\{e_1, e_2, \ldots, e_n\}$.

Para obtermos uma sequência ortonormal $\{e_1^*, e_2^*, \ldots, e_n^*\}$, basta fazer:

$$e_i^* = \frac{e_i}{\| e_i \|}, \quad i = 1, 2, \ldots, n.$$

Teorema 1.6 - *Todo espaço euclidiano n dimensional tem uma base ortogonal e uma base ortonormal.*

Prova: Todo espaço euclidiano E é um espaço vetorial, e, portanto tem uma base. Seja $f_1, f_2, \ldots, f_n$ uma base desse espaço euclidiano. Vamos construir a partir de $f_1, f_2, \ldots, f_n$ uma base ortogonal de E. Seja $\{e_1, e_2, \ldots, e_n\}$ a base procurada.

Tomamos e_1 como sendo igual ao primeiro elemento da sequência dada, isto é:

$$e_1 = f_1 .$$

O elemento e_2 será tomado como combinação linear do segundo elemento da sequência dada e e_1, ou seja:

$$e_2 = f_2 + \alpha_1 e_1 ,$$

onde α_1 é escolhido de tal maneira que e_2 seja ortogonal a e_1. Assim: $(e_2, e_1) = 0 \rightarrow (f_2 + \alpha_1 e_1, e_1) = 0$. Portanto, segue que:

$$\alpha_1 = -\frac{(f_2, e_1)}{(e_1, e_1)} .$$

Vamos supor que já temos construído os vetores: $e_1, e_2, \ldots, e_{k-1}$, dois a dois ortogonais. O elemento e_k será tomado como combinação linear do k^o elemento da sequência dada e todos os e_i, já calculados, isto é:

$$e_k = f_k + \alpha_{k-1} e_{k-1} + \alpha_{k-2} e_{k-2} + \ldots + \alpha_1 e_1 ,$$

onde os $\alpha_i, \; i = 1, 2, \ldots, k-1$, são determinados de tal maneira que e_k seja ortogonal a todos os e_i já calculados. Assim, devemos ter: $(e_k, e_i) = 0, \; i \; = \; 1, 2, \ldots, k-1$, ou seja:

$$\begin{array}{rcl} (e_k, e_1) & = & (f_k + \alpha_{k-1} e_{k-1} + \ldots + \alpha_1 e_1, e_1) \; = \; 0 \; , \\ (e_k, e_2) & = & (f_k + \alpha_{k-1} e_{k-1} + \ldots + \alpha_1 e_1, e_2) \; = \; 0 \; , \\ \vdots & & \\ (e_k, e_{k-1}) & = & (f_k + \alpha_{k-1} e_{k-1} + \ldots + \alpha_1 e_1, e_{k-1}) \; = \; 0 \; . \end{array}$$

Desde que os vetores $e_1, e_2, \ldots, e_{k-1}$ foram construídos dois a dois ortogonais, obtemos:

$$\begin{array}{rcl} (f_k, e_1) \; + \; \alpha_1 \, (e_1, e_1) & = & 0 \; , \\ (f_k, e_2) \; + \; \alpha_2 \, (e_2, e_2) & = & 0 \; , \\ \vdots & & \\ (f_k, e_{k-1}) \; + \; \alpha_{k-1} \, (e_{k-1}, e_{k-1}) & = & 0 \; . \end{array}$$

Portanto, segue que:

$$\begin{array}{rcl} \alpha_1 & = & - \dfrac{(f_k, e_1)}{(e_1, e_1} \; , \\ \alpha_2 & = & - \dfrac{(f_k, e_2)}{(e_2, e_2)} \; , \\ \vdots & & \\ \alpha_{k-1} & = & - \dfrac{(f_k, e_{k-1})}{(e_{k-1}, e_{k-1})} \; . \end{array}$$

Mostremos agora que $e_k \neq 0$. De fato, temos que e_k é combinação linear dos vetores $e_1, e_2, \ldots, e_{k-1}, f_k$. Mas e_{k-1} pode ser escrito com combinação linear dos vetores $e_1, e_2, \ldots, e_{k-2}, f_{k-1}$ e assim por diante. Então, substituindo, teremos:

$$e_k \; = \; a_1 \; f_1 \; + \; a_2 \; f_2 \; + \ldots + \; a_{k-1} \; f_{k-1} \; + \; f_k \; ,$$

e como $f_1, f_2, \ldots, f_k$, são linearmente independentes, temos que $e_k \neq 0$; qualquer que seja k.

Assim, usando $e_1, e_2, \ldots, e_{k-1}$ e f_k construímos e_k. Analogamente com $e_1, e_2, \ldots, e_k$ e f_{k+1} construímos e_{k+1}. Continuando o processo, construímos os n vetores dois a dois ortogonais. Assim esses vetores formam uma base ortogonal de E. Tomando:

$$e_i^* \; = \; \frac{e_i}{\parallel e_i \parallel} \; , \quad i = 1, 2, \ldots, n \; ;$$

teremos uma base ortonormal de E.

Chama-se **processo de Gram-Schmidt** a construção passo a passo (descrita na prova do teorema 1.6) para converter uma base arbitrária em base ortogonal.

Exemplo 1.15 - *Construir a partir de*

$$f_1 = (1, \; -2, \; 0)^t \; , f_2 = (0, \; 1, \; 1)^t \; , f_3 = (1, \; 0, \; -1)^t;$$

uma sequência de vetores ortonormais e_1^*, e_2^*, e_3^**, relativamente ao produto escalar usual do* $I\!R^3$*, usando o processo de Gram-Schmidt.*

Solução: Temos:

$$\begin{aligned}
e_1 &= f_1 = (1,\ -2,\ 0)^t \ . \\
e_2 &= f_2 + \alpha_1\, e_1, \quad \text{onde} \\
&\quad \alpha_1 = -\frac{(f_2, e_1)}{(e_1, e_1)} = -\frac{-2}{5} = \frac{2}{5} \quad \Rightarrow \\
e_2 &= (0,\ 1,\ 1)^t + \frac{2}{5}\,(1,\ -2,\ 0)^t = \left(\frac{2}{5}\ ,\ \frac{1}{5}\ ,\ 1\right)^t . \\
e_3 &= f_3 + \alpha_2 e_2 + \alpha_1 e_1, \quad \text{onde} \\
&\quad \alpha_2 = -\frac{(f_3, e_2)}{(e_2, e_2)} = -\frac{-3/5}{6/5} = \frac{1}{2}\ , \\
&\quad \alpha_1 = -\frac{(f_3, e_1)}{(e_1, e_1)} = -\frac{1}{5} \quad \Rightarrow \\
e_3 &= (1,\ 0,\ -1)^t + \frac{1}{2}\left(\frac{2}{5},\ \frac{1}{5},\ 1\right) - \frac{1}{5}\,(1,\ -2,\ 0)^t = \left(1,\ \frac{1}{2},\ -\frac{1}{2}\right)^t .
\end{aligned}$$

Assim e_1, e_2, e_3 são dois a dois ortogonais. Para obtermos a sequência ortonormal e_1^*, e_2^*, e_3^*, fazemos:

$$\begin{aligned}
e_1^* &= \frac{e_1}{\parallel e_1 \parallel} = \frac{e_1}{\sqrt{(e_1, e_1)}} = \frac{(1,\ -2,\ 0)^t}{\sqrt{1^2 + (-2)^2 + 0^2}} = \left(\frac{1}{\sqrt{5}},\ \frac{-2}{\sqrt{5}},\ 0\right)^t ; \\
e_2^* &= \frac{e_2}{\parallel e_2 \parallel} = \frac{e_2}{\sqrt{(e_2, e_2)}} = \frac{(2/5,\ 1/5,\ 1)^t}{\sqrt{(2/5)^2 + (1/5)^2 + 1^2}} = \sqrt{\frac{5}{6}}\ \left(\frac{2}{5},\ \frac{1}{5},\ 1\right)^t ; \\
e_3^* &= \frac{e_3}{\parallel e_3 \parallel} = \frac{e_3}{\sqrt{(e_3, e_3)}} = \frac{(1,\ 1/2,\ -1/2)^t}{\sqrt{1^2 + (1/2)^2 + (-1/2)^2}} = \sqrt{\frac{2}{3}}\ \left(1,\ \frac{1}{2},\ -12\right)^t .
\end{aligned}$$

Exemplo 1.16 - *Dada a sequência de polinômios independentes* $\{1,\ x,\ x^2\}$ *obter, no intervalo* $[-1, 1]$, *uma sequência ortogonal de polinômios* $\{P_0(x), P_1(x), P_2(x)\}$ *relativamente ao produto escalar* $(f, g) = \int_{-1}^{1} f(x)\ g(x)\ dx$.

Solução: Temos:

$$\begin{aligned}
P_0(x) &= 1\ , \\
P_1(x) &= x + \alpha_0 P_0(x)\ , \quad \text{onde} \\
\alpha_0 &= -\frac{(x,\ P_0(x))}{(P_0(x), P_0(x))} = -\frac{\int_{-1}^{1} x\ dx}{\int_{-1}^{1}\ dx} = \left.\frac{x^2/2}{x}\right]_{-1}^{1} = 0 \quad \Rightarrow \\
P_1(x) &= x + 0 \times 1 = x.
\end{aligned}$$

$$\begin{aligned}
P_2(x) &= x^2 + \alpha_1 P_1(x) + \alpha_0 P_0(x), \quad \text{onde} \\
\alpha_1 &= -\frac{(x^2,\ P_1(x))}{(P_1(x), P_1(x))} = -\frac{\int_{-1}^{1}\ x^3\ dx}{\int_{-1}^{1}\ x^2\ dx} = \left.\frac{x^4/4}{x^3/3}\right]_{-1}^{1} = 0\ , \\
\alpha_0 &= -\frac{(x^2,\ P_0(x))}{(P_0(x), P_0(x))} = -\frac{\int_{-1}^{1}\ x^2\ dx}{\int_{-1}^{1}\ dx} = -\left.\frac{x^3/3}{x}\right]_{-1}^{1} = -\frac{2/3}{2} = -\frac{1}{3} \quad \Rightarrow \\
P_2(x) &= x^2 + 0 \times x - \frac{1}{3} \times 1 = x^2 - \frac{1}{3}.
\end{aligned}$$

Assim $P_0(x), P_1(x), P_2(x)$ são dois a dois ortogonais.

Observe que sempre que desejarmos obter uma sequência de polinômios ortogonais sobre um determinado intervalo, podemos tomar a sequência $1,\ x,\ x^2, \ldots$ como sendo a sequência original e ortogonalizá-la.

Exercícios

1.19 - *Usando o processo de Gram-Schmidt e o produto escalar usual do $I\!R^3$, ortonormalizar a base:*

$$e_1 = (1,\ 1,\ 1)^t\ , e_2 = (1,\ -1,\ 1)^t\ , e_3 = (-1,\ 0,\ 1)^t\ .$$

1.20 - *Os vetores $\{(0,\ 2,\ 1,\ 0)^t, (1,\ -1,\ 0,\ 0)^t, (1,\ 2,\ 0,\ -1)^t, (1,\ 0,\ 0,\ 1)^t\}$ constituem uma base não ortonormal do $I\!R^4$. Construir a partir desses vetores, uma base ortonormal para o $I\!R^4$, usando o processo de Gram-Schmidt.*

1.21 - *Ortonormalize a sequência de polinômios obtida no exemplo 1.16.*

1.22 - *Usando o produto escalar usual em $C[1,2]$ e o processo de Gram-Schmidt construa uma sequência de polinômios ortonormais.*

1.4 Projeção Ortogonal

Veremos aqui a projeção ortogonal de um vetor sobre outro bem como a projeção ortogonal de um vetor sobre um sub-espaço. Esse último será utilizado no estudo de aproximações de funções pelo método dos mínimos quadrados.

Projeção Ortogonal de um Vetor sobre Outro

Sejam x e y vetores não nulos. Escolhemos um número real λ tal que $\lambda\ y$ seja ortogonal a $x - \lambda\ y$, como sugere a Figura 1.2, no caso em que $E = I\!R^2$.

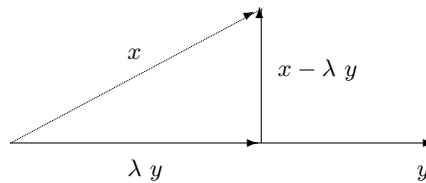



Figura 1.2

De $\lambda\ y \perp (x - \lambda\ y)$, concluímos que $(\lambda\ y, x - \lambda\ y) = 0$. Portanto, aplicando P_3, segue que:

$$\lambda(y,x) - \lambda^2(y,y) \ = \ 0 \ \rightarrow \ \lambda \ = \ \frac{(x,y)}{(y,y)}\ .$$

Assim, obtemos a seguinte definição.

Definição 1.14 - *Num espaço euclidiano real, chama-se* **projeção ortogonal** *de x sobre $y, y \neq \theta$, o vetor z definido por:*

$$z \ = \quad (\textit{projeção de } x \textit{ sobre } y) \ = \ \frac{(x,y)}{(y,y)}\ y.$$

Se $\| y \| = 1$, então a projeção de x sobre y é dada por $(x, y)\ y$.

Projeção Ortogonal de um Vetor sobre um Sub-Espaço

Seja E um espaço euclidiano e seja E', de dimensão finita n, um sub-espaço de E.

Seja v um vetor de E não pertencente a E'.

O problema que desejamos resolver agora é o de obter um vetor $v_0 \in E'$ tal que $v - v_0$ seja ortogonal a todo vetor de E'. (A Figura 1.3 ilustra o problema, para o caso em que $E = I\!R^3$ e $E' = I\!R^2$).

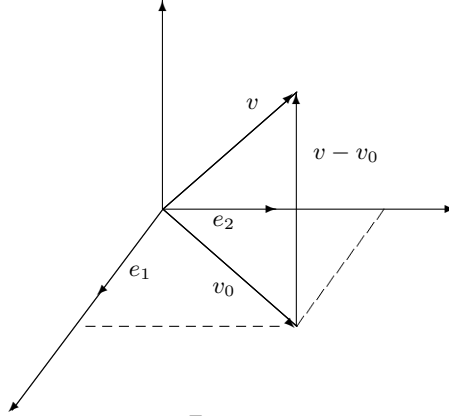



Figura 1.3

Seja $\{e_1, e_2, \ldots, e_n\}$ uma base de E'. Como $v_0 \in E'$, v_0 pode ser escrito como combinação linear dos vetores da base de E', isto é:

$$v_0 \ = \ \gamma_1\ e_1 \ + \ \gamma_2\ e_2 \ + \ldots + \ \gamma_n\ e_n\ . \tag{1.17}$$

O nosso problema consiste em determinar, caso possível, as coordenadas $\gamma_1, \gamma_2, \ldots, \gamma_n$ de v_0.

Sabemos que se $v - v_0$ deve ser ortogonal a todo vetor de E' então é necessário e suficiente que $v - v_0$ seja ortogonal a todo vetor de uma base de E' (Teorema 1.2). Então, devemos ter:

$$\begin{array}{l} (v - v_0, e_j) \ = \ 0 \quad \text{para} \quad j = 1, 2, \ldots, n \ \ ; \ \ \text{ou seja :} \\ (v - (\gamma_1\ e_1 + \gamma_2\ e_2 + \ldots + \gamma_n\ e_n)\, , e_j) \ = \ 0\ ,\ j = 1, 2, \ldots, n. \end{array}$$

A aplicação de P_2 e P_3, fornece:

$$\gamma_1\,(e_1, e_j) \ + \ \gamma_2\,(e_2, e_j) \ + \ldots + \ \gamma_n\,(e_n, e_j) \ = \ (v, e_j)\ \ ,\ j = 1, \ldots, n\ .$$

Tais equações são conhecidas por **equações normais**.

Assim, para obtermos as coordenadas de v_0 na base $\{e_1, e_2, \ldots, e_n\}$, devemos resolver o sistema de equações lineares:

$$\begin{pmatrix} (e_1, e_1) & (e_2, e_1) & \ldots & (e_n, e_1) \\ (e_1, e_2) & (e_2, e_2) & \ldots & (e_n, e_2) \\ \ldots & & & \\ (e_1, e_n) & (e_2, e_n) & \ldots & (e_n, e_n) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_n \end{pmatrix} = \begin{pmatrix} (v, e_1) \\ (v, e_2) \\ \vdots \\ (v, e_n) \end{pmatrix}, \tag{1.18}$$

cuja matriz dos coeficientes é simétrica.

Mostremos agora que o sistema (1.18) tem uma e uma só solução, isto é, que o problema de determinação do vetor $v_0 \in E'$, tal que $v - v_0$ seja ortogonal a todo vetor de E', tem solução única.

O vetor v_0 é denominado **projeção ortogonal** de v sobre o sub-espaço E'.

Vamos supor que nossa base de partida fosse uma base $\{e'_1, e'_2, \ldots, e'_n\}$ ortonormal. Esta não seria uma hipótese restritiva, uma vez que é sempre possível passar-se de uma dada base para uma base ortonormal, (ver processo de Gram-Schmidt).

Em termos da base ortonormal considerada o vetor v_0 se exprimiria como:

$$v_0 \; = \; \gamma'_1 \; e'_1 \; + \; \gamma'_2 \; e'_2 + \ldots + \gamma'_n \; e'_n .$$

O sistema linear (1.18) se reduziria a:

$$\begin{pmatrix} 1 & & & \bigcirc \\ & 1 & & \\ & & \ddots & \\ \bigcirc & & & 1 \end{pmatrix} \begin{pmatrix} \gamma'_1 \\ \gamma'_2 \\ \vdots \\ \gamma'_n \end{pmatrix} = \begin{pmatrix} (v, e'_1) \\ (v, e'_2) \\ \vdots \\ (v, e'_n) \end{pmatrix} ,$$

ou simplesmente a:

$$\gamma'_j \; = \; \left(v, e'_j \right) , \quad j \; = \; 1, 2, \ldots, n \; , \tag{1.19}$$

e portanto os γ'_j seriam univocamente determinados.

Sabemos que, conhecidas as coordenandas de um vetor numa base, suas coordenadas em outra qualquer base são também univocamente determinadas. Assim, o sistema (1.18) tem uma única solução $(\gamma_1, \gamma_2, \ldots, \gamma_n)^t$ e a matriz do sistema em apreço é sempre não singular. A projeção ortogonal v_0 de v sobre E' é, portanto, única.

Exemplo 1.17 - *Seja $E = C[-1,1]$, com $(f,g) = \int_{-1}^{1} f(x)g(x)dx$. Seja $K_2(x)$ o sub-espaço dos polinômios de grau ≤ 2. O conjunto $\{L_0(x) = 1, \; L_1(x) = x, \; L_2(x) = x^2\}$ constitui uma base de $K_2(x)$. Determinar a projeção ortogonal de $f(x) = \dfrac{1}{x+4}$ sobre $k_2(x)$.*

Solução: De (1.17) temos: $f_0 = \gamma_0 \; L_0(x) + \gamma_1 \; L_1(x) + \gamma_2 \; L_2(x)$. Assim, devemos determinar $\gamma_0, \gamma_1, \gamma_2$. Para tanto, montamos o sistema (1.18):

$$\begin{pmatrix} (L_0, L_0) & (L_1, L_0) & (L_2, L_0) \\ (L_0, L_1) & (L_1, L_1) & (L_2, L_1) \\ (L_0, L_2) & (L_1, L_2) & (L_2, L_2) \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} (f, L_0) \\ (f, L_1) \\ (f, L_2) \end{pmatrix} ;$$

onde:

$$\begin{aligned} (L_0, L_0) &= \int_{-1}^{1} dx \; = \; x]_{-1}^{1} \; = \; 2 \; , \\ (L_1, L_0) &= (L_0, L_1) \; = \; \int_{-1}^{1} x \; dx \; = \; \left. \frac{x^2}{2} \right]_{-1}^{1} \; = \; 0 \; , \\ (L_2, L_0) &= (L_0, L_2) \; = \; \int_{-1}^{1} x^2 \; dx \; = \; \left. \frac{x^3}{3} \right]_{-1}^{1} \; = \; \frac{2}{3} \; , \\ (L_1, L_1) &= \int_{-1}^{1} x^2 dx \; = \; \frac{2}{3} \; , \end{aligned}$$

$$
\begin{aligned}
(L_2, L_1) &= (L_1, L_2) = \int_{-1}^{1} x^3 \, dx = \frac{x^4}{4}\Big]_{-1}^{1} = 0 \, , \\
(L_2, L_2) &= \int_{-1}^{1} x^4 \, dx = \frac{x^5}{5}\Big]_{-1}^{1} = \frac{2}{5} \, , \\
(f, L_0) &= \int_{-1}^{1} \frac{1}{x+4} \, dx = (ln\ (x+4))]_{-1}^{1} = 0.51083 \, , \\
(f, L_1) &= \int_{-1}^{1} \frac{x}{x+4} \, dx = \int_{-1}^{1} \left(1 - \frac{4}{x+4}\right) dx \\
&= (x - 4\ ln\ (x+4))]_{-1}^{1} = -0.04332 \, , \\
(f, L_2) &= \int_{-1}^{1} \frac{x^2}{x+4} dx = \int_{-1}^{1} \left(x - 4 + \frac{16}{x+4}\right) dx \\
&= \left(\frac{x^2}{2} - 4\,x + 16\ ln\ (x+4)\right)\Big]_{-1}^{1} = 0.17328 \, .
\end{aligned}
$$

Assim, obtemos o sistema linear:

$$
\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix}
\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}
=
\begin{pmatrix} 0.51083 \\ -0.04332 \\ 0.17328 \end{pmatrix} \, ,
$$

cuja solução é: $\gamma_0 = 0.24979$; $\gamma_1 = -\ 0.06498$; $\gamma_2 = 0.01688$. Então, a projeção ortogonal de $f(x) = \dfrac{1}{x+4}$ sobre $K_2(x)$ é:

$$
\begin{aligned}
f_0 &= 0.24979\ L_0(x) \ - \ 0.06498\ L_1(x) \ + \ 0.01688\ L_2(x) \\
&= 0.24979 \ - \ 0.06498\ x \ + \ 0.01688\ x^2.
\end{aligned}
$$

Teorema 1.7 - **Teorema da Melhor Aproximação** - *Seja E' um sub-espaço de dimensão finita de um espaço euclidiano E. Se v for um vetor pertencente a E, então v_0, a projeção ortogonal de v sobre E', será a* **melhor aproximação** *para v no sentido de que*

$$
\| \, v - v_0 \, \| < \| \, v - y \, \| \, , \tag{1.20}
$$

para qualquer que seja $y \in E'$, tal que $y \neq v_0$.

Prova: Devemos mostrar que a menor distância de v ao sub-espaço E' é a distância entre v e o pé da perpendicular traçada da extremidade de v sobre E'. (A Figura 1.3 ilustra o problema para o caso em que $E = I\!R^3$ e $E' = I\!R^2$).

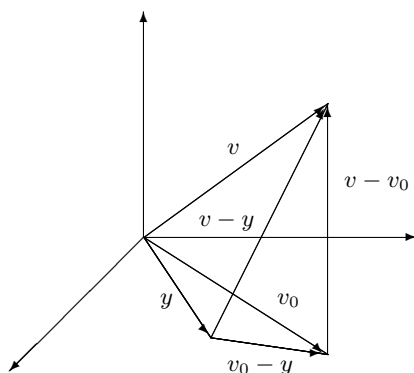



Figura 1.4

Como $y,\ v_0 \in E'$ também $v_0 - y \in E'$ e é portanto ortogonal a $v - v_0$. Assim, obtemos, sucessivamente:

$$\begin{aligned}(v-y, v-y) &= (v - y + v_0 - v_0, v - y + v_0 - v_0) \\ &= (v - v_0, v - v_0) + 2\,(v - v_0, v_0 - y) \ + \ (v_0 - y, v_0 - y)\,.\end{aligned}$$

Portanto:

$$\| \ v - y \ \|^2 = \| \ v - v_0 \ \|^2 \ + \ \| \ v_0 - y \ \|^2 \ . \qquad (1.21)$$

Como, por hipótese, $y \neq v_0$, concluímos que $\| \ v_0 - y \ \| > \ 0$. Daí, e da igualdade (1.21), obtemos, finalmente:

$$\| \ v - y \ \| > \| \ v - v_0 \ \| \ .$$

Assim, a desigualdade (1.20) mostra que a projeção ortogonal v_0 de v sobre E' é tal que a menor distância de v sobre E' é a distância de v a v_0.

Exercícios

1.23 - *Seja* $x = (1,\ 7,\ 10)^t$ *um vetor do* $I\!R^3$ *em relação à base canônica. Considere o sub-espaço* E' *do* $I\!R^3$*, gerado pelos vetores* $f_1 = (1,\ 1,\ 0)^t$ *e* $f_2 = (0,\ 1,\ 1)^t$*. Determine a projeção ortogonal de* x *sobre* E'*.*

1.24 - *Seja* $E = C[0,1]$*, com* $(f,g) = \int_0^1 f(x)g(x)dx$*. Seja* $K_2(x)$ *o sub-espaço dos polinômios de grau* ≤ 2*. O conjunto* $\{Q_0(x) = 3,\ Q_1(x) = x - 3,\ Q_2(x) = x^2 - x\}$ *constitui uma base de* $K_2(x)$*. Determinar a projeção ortogonal de* $f(x) = \dfrac{1}{x^4}$ *sobre* $k_2(x)$*.*

1.5 Auto-Valores e Auto-Vetores

Nessa seção, investigaremos a teoria de um operador linear T num K-espaço vetorial V de dimensão finita. Também associaremos um polinômio ao operador T: seu polinômio característico. Esse polinômio e suas raízes desempenham papel proeminente na investigação de T. Apresentaremos também alguns conceitos que serão de grande utilidade na obtenção de métodos para determinação numérica de auto-valores e auto-vetores de matrizes.

Definição 1.15 - *Uma* **transformação linear** T *de um* K*-espaço vetorial* V *em um* K*-espaço vetorial* U*,* $T : V \to U$*, é uma correspondência que associa a cada vetor* x *de* V *um vetor* $T(x)$ *em* U *de modo que:*

$$T(\alpha x + \beta y) \ = \ \alpha T(x) + \beta T(y) \ , \forall x,\ y \in V, \forall \alpha,\ \beta \in K.$$

Em particular, se $U = V$, então dizemos que T é um **operador linear** num K-espaço vetorial V.

Definição 1.16 - *Um escalar $\lambda \in K$ é um* **auto-valor** *de T se existe um vetor* **não nulo** *$v \in V$ tal que:*

$$T(v) = \lambda \ v \ .$$

Todo vetor v satisfazendo essa relação é um **auto-vetor** *de T correspondente ao auto-valor λ.*

Observações:

1. Se λ é um auto-valor de T, então o operador linear pode apenas variar o módulo e o sentido do vetor, nunca sua direção.
2. Os termos valor característico e vetor característico (ou valor próprio e vetor próprio) são frequentemente usados ao invés de auto-valor e auto-vetor.

Daremos a seguir alguns exemplos.

Exemplo 1.18 - *Seja $I : V \to V$ o operador identidade onde $V = I\!R^n$. Determinar seus auto-valores e auto-vetores.*

Solução: Para cada $v \in V$, temos que:

$$I(v) \ = \ v \ = \ 1 \cdot v \ .$$

Portanto, **1** é auto-valor de I e todo vetor não nulo em V é um auto-vetor correspondente ao auto-valor 1.

Exemplo 1.19 - *Seja $D : V \to V$ o operador diferencial onde V é o espaço vetorial das funções diferenciáveis. Determinar um auto-valor de D e seu correspondente auto-vetor.*

Solução: Temos que $e^{kt} \in V$, e, sabemos que:

$$D\left(e^{kt}\right) = k \ e^{kt} \ .$$

Logo, **k** é um auto-valor de D e e^{kt} é auto-vetor de D correspondente ao auto-valor k.

Exemplo 1.20 - *Seja $T : \ I\!R^2 \to I\!R^2$ o operador linear que gira cada vetor $v \in I\!R^2$ de um ângulo ψ. Determinar os auto-valores e correspondentes auto-vetores nos seguintes casos:*

$$a)\psi \ = \ 2n\pi \ , \qquad b)\psi \ = \ (2n+1)\pi \ , \qquad c)\psi \ = \ \left(\frac{2n+1}{2}\right)\pi \ .$$

Solução: Temos que o operador linear que gira cada vetor de um ângulo ψ é dado por uma matriz chamada **matriz de rotação**. No caso em que $V = I\!R^2$ essa matriz é dada por:

$$T \ = \ \begin{pmatrix} cos \ \psi & sen \ \psi \\ -sen \ \psi & cos\psi \end{pmatrix} \ .$$

Seja $v \in I\!R^2$, então $v = (v_1, \ v_2)^t$. Podemos considerar nos três casos $n = 1$, visto que para valores maiores de n teremos apenas um número maior de rotações. Assim, para:

a) $\psi \ = \ 2\pi$, temos:

$$\begin{pmatrix} cos \ 2\pi & sen \ 2\pi \\ -sen \ 2\pi & cos \ 2\pi \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \ = \ \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \ = \ 1 \ \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \ ,$$

b) $\psi = 3\pi$, temos:

$$\begin{pmatrix} cos\ 3\pi & sen\ 3\pi \\ -sen\ 3\pi & cos\ 3\pi \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -v_1 \\ -v_2 \end{pmatrix} = -1 \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} ,$$

c) $\psi = \frac{3\pi}{2}$

$$\begin{pmatrix} cos\ \frac{3\pi}{2} & sen\ \frac{3\pi}{2} \\ -sen\ \frac{3\pi}{2} & cos\ \frac{3\pi}{2} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -v_2 \\ v_1 \end{pmatrix} \neq \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} .$$

Logo, os auto-valores de T são:

$$\mathbf{1} \text{ se } \psi = 2n\pi \ , \quad \mathbf{-1} \text{ se } \psi = (2n+1)\pi \ ,$$

e em ambos os casos todo vetor não nulo do $I\!R^2$ é auto-vetor de T. Se $\psi = (\frac{2n+1}{2})\pi$, T não tem auto-valores e portanto T não tem auto-vetores. Observe que neste caso o operador linear está variando a direção do vetor.

Se A é uma matriz quadrada $n \times n$ sobre K, então um auto-valor de A significa um auto-valor de A encarado como operador em K^n. Isto é, $\lambda \in K$ é um auto-valor de A se, para algum vetor (coluna) não nulo $v \in K^n, Av = \lambda v$. Nesse caso, v é um auto-vetor de A correspondente a λ.

Exemplo 1.21 - *Seja:*

$$A = \begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix} .$$

Determinar os auto-valores e auto-vetores de A.

Solução: Procuramos um escalar λ e um vetor não nulo $v = (v_1, \ v_2)^t$ tais que $Av = \lambda v$. Assim:

$$\begin{pmatrix} 3 & 4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} .$$

A equação matricial acima é equivalente ao sistema homogêneo:

$$\begin{cases} 3v_1 + 4v_2 = \lambda v_1 \\ 2v_1 + v_2 = \lambda v_2 \end{cases} \quad \text{ou} \quad \begin{cases} (3-\lambda)v_1 + 4v_2 = 0 \\ 2v_1 + (1-\lambda)v_2 = 0 \end{cases} \tag{1.22}$$

Para que o sistema homogêneo tenha solução não nula, o determinante da matriz dos coeficientes deve ser igual a zero. Logo:

$$\begin{vmatrix} (3-\lambda) & 4 \\ 2 & (1-\lambda) \end{vmatrix} = \lambda^2 - 4\lambda - 5 = (\lambda-5)(\lambda+1) = 0 .$$

Assim, λ é um auto-valor de A se e somente se, $\lambda = 5$ ou $\lambda = -1$.
Fazendo $\lambda = 5$ em (1.22), obtemos:

$$\begin{cases} -2v_1 + 4v_2 = 0 \\ 2v_1 - 4v_2 = 0 \end{cases}$$

ou simplesmente, $v_1 - 2v_2 = 0 \Rightarrow v_1 = 2v_2$. Assim $v = (v_1, \ v_2)^t = (2, \ 1)^t$ é um auto-vetor correspondente ao auto-valor $\lambda = 5$. Qualquer outro auto-vetor correspondente a $\lambda = 5$ é um múltiplo de v.

Fazendo $\lambda = -1$ em (1.22), obtemos:

$$\begin{cases} 4v_1 + 4v_2 = 0 \\ 2v_1 + 2v_2 = 0 \end{cases}$$

ou simplesmente, $v_1 + v_2 = 0 \Rightarrow v_1 = -v_2$. Assim $v = (v_1,\ v_2)^t = (1,\ -1)^t$ é um auto-vetor correspondente ao auto-valor $\lambda = -1$ e novamente, qualquer outro auto-vetor correspondente a $\lambda - 1$ é um múltiplo de v.

Definição 1.17 - *Dada uma matriz quadrada $A, n \times n$, a matriz:*

$$A - \lambda I = \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{pmatrix},$$

onde I é a matriz identidade de ordem n e λ é um parâmetro, é chamada **matriz característica** *de A. Seu determinante , $|A - \lambda I|$, é um polinômio de grau n em λ chamado* **polinômio característico** *de A.*

Exemplo 1.22 - *Seja $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. Determinar seu polinômio característico.*

Solução: Para calcular o polinômio característico de A, basta calcular o determinante de $A - \lambda I$. Assim:

$$|A - \lambda I| = \begin{vmatrix} 1 - \lambda & 2 \\ 3 & 4 - \lambda \end{vmatrix} = \underbrace{\lambda^2 - 5\lambda - 2\ .}_{\text{polinômio característico.}}$$

Exercícios

1.25 - *Prove que os auto-valores de A são os zeros do polinômio característico.*

1.26 - *Prove que: se $\lambda_1,\ \lambda_2,\ \dots,\ \lambda_n$ são auto-valores de A então $\lambda_1^k,\ \lambda_2^k,\ \dots,\ \lambda_n^k$ são auto-valores de A^k.*

Como já dissemos anteriomente estudaremos, (no Capítulo 7), métodos numéricos para determinação de auto-valores e auto-vetores de matrizes. Tais métodos para serem obtidos dependem de alguns conceitos os quais passamos a discutir agora.

Polinômio de Matrizes

Definição 1.18 *Seja:*

$$P(t) = a_0\ t^n + a_1\ t^{n-1} + \dots + a_{n-1}\ t + a_n\ ,$$

um polinômio de grau n onde os $a_i,\ i = 1, 2, \dots, n$ são reais.

Se A é uma matriz quadrada real, então definimos:

$$P(A) = a_0\ A^n + a_1\ A^{n-1} + \dots + a_{n-1}\ A + a_n\ I\ ,$$

como sendo o **polinômio da matriz A**. *Na expressão acima I é a matriz identidade.*

Em particular, se $P(A) = \theta$, (matriz nula), dizemos que A é um zero de $P(t)$.

Exemplo 1.23 - *Seja* $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$. *Calcular* $P(A)$ *e* $Q(A)$, *sabendo que:* $P(t) = 2t^3 - 3t + 7$ *e* $Q(t) = t^2 - 5t - 2$.

Solução: Temos que:

$$P(A) \;=\; 2 \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^3 - 3 \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + 7 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 18 & 14 \\ 21 & 39 \end{pmatrix},$$

e

$$Q(A) \;=\; \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^2 - 5 \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - 2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Assim, A é um zero de $Q(t)$. Note que $Q(t)$ é o polinômio característico de A.

Teorema 1.8 - *(Teorema de Cayley-Hamilton) - Toda matriz é um zero do seu polinômio característico.*

Prova: A prova desse teorema pode ser encontrada em [Barnett, 1990].

Transformações de Similaridades (ou Semelhança)

Existem métodos numéricos que determinam todos os auto-valores de uma matriz sem determinar a expressão do polinômio característico. Tais métodos são obtidos usando-se transformações de similaridade.

Definição 1.19 - *Uma matriz B é* **similar** *(ou* **semelhante***) a uma matriz A se $\exists$ uma matriz C não singular tal que:*

$$B = C^{-1}AC\ ,$$

e dizemos que B foi obtida de A por transformação de semelhança.

Teorema 1.9 - *Sejam A e B matrizes similares. Então:*

i) *A e B possuem os mesmos auto-valores.*

ii) *Se v é auto-vetor de A associado a λ, então $C^{-1}v$ é auto-vetor de $B = C^{-1}AC$ associado a λ.*

Prova: Seja $B = C^{-1}AC$, e suponha que λ é auto-valor de A e v seu correspondente auto-vetor. Temos então, que $det(A - \lambda I)$ é o polinômio característico de A.

i) Temos :

$$\begin{aligned} det(B - \lambda I) &= det(C^{-1}AC - \lambda I) \\ &= det(C^{-1}(A - \lambda I)C) \\ &= detC^{-1}det(A - \lambda I)detC \\ &= det(A - \lambda I)det(\underbrace{C^{-1}C}_{=I}) = det(A - \lambda I)\ . \end{aligned}$$

Portanto A e B possuem o mesmo polinômio característico. Logo λ é auto-valor de B.

ii) Agora $Av = \lambda v$ e desde que $B = C^{-1}AC \;\Rightarrow A = CBC^{-1}$. Portanto $CBC^{-1}v = \lambda v$. Assim:

$$BC^{-1}v = C^{-1}\lambda v = \lambda C^{-1}v\ .$$

Portanto $B(C^{-1}v) = \lambda(C^{-1}v)$. Logo $C^{-1}v$ é auto-vetor de B associado ao auto-valor λ.

Lema 1.1 - *Seja A uma matriz de ordem n com auto-valores λ_i e correspondentes auto-vetores v_i, os quais vamos supor sejam linearmente independentes, e seja*

$$D = \begin{pmatrix} \lambda_1 & & & & \bigcirc \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ \bigcirc & & & & \lambda_n \end{pmatrix} .$$

Então $D = V^{-1}AV$ se e somente a i-ésima coluna de V é v_i.

Prova: Se a i-ésima coluna de V é denotada por v_i então a i-ésima coluna de AV e VD são, Av_i e $\lambda_i v_i$, respectivamente. Portanto os vetores v_i são os auto-vetores de A se e somente se $AV = VD$. Esta equação pode ser rearranjada como: $V^{-1}AV$ desde que V seja inversível, e este é o caso pois as colunas de V são linearmente independentes.

Matriz de Rotação e Matriz Ortogonal

Alguns métodos numéricos são obtidos usando-se matrizes que possuem características especiais. Assim, passamos a descrever tais matrizes.

No $I\!R^2$ as matrizes:

$$\begin{pmatrix} cos\ \varphi & sen\ \varphi \\ -sen\ \varphi & cos\ \varphi \end{pmatrix} , \quad \begin{pmatrix} cos\ \varphi & -sen\ \varphi \\ sen\ \varphi & cos\ \varphi \end{pmatrix} ,$$

rotacionam cada vetor do $I\!R^2$, no sentido horário e anti-horário, respectivamente, de um ângulo φ, e porisso são chamadas de **Matrizes de Rotação**.

No $I\!R^3$ a matriz:

$$\begin{pmatrix} cos\ \varphi & 0 & sen\ \varphi \\ 0 & 1 & 0 \\ -sen\ \varphi & 0 & cos\ \varphi \end{pmatrix} ,$$

é uma matriz de rotação, no sentido horário, de um ângulo φ no plano x, z.

No $I\!R^n$ a matriz:

$$U = \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & cos\ \varphi & 0 & \dots & 0 & sen\ \varphi & & & \\ & & & & 1 & & & & & & \\ & & & & \vdots & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & -sen\ \varphi & 0 & \dots & 0 & cos\ \varphi & & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix} \tag{1.23}$$

onde:

$$\begin{cases} u_{pp} = u_{qq} = cos\varphi \\ u_{pg} = -u_{qp} = sen\varphi \\ u_{ij} = 1, i \neq p, i \neq q \\ uij = 0, \text{ no resto} \end{cases}$$

é uma matriz de rotação de um ângulo φ no plano dos eixos p e q.

Uma **Matriz Ortogonal** U é caracterizada por:

$$U^t U = U U^t = I \ ,$$

onde I: matriz identidade. Portanto $U^t = U^{-1}$.

Observe que matrizes de rotação são matrizes ortogonais.

Propriedades de Matrizes Ortogonais

1) As linhas de U satisfazem:

$$\sum_{j=1}^{n} (u_{ij})^2 = 1 \text{ (produto de uma linha por ela mesma)} \ ,$$

$$\sum_{\substack{j=1 \\ i \neq k}}^{n} u_{ij} \ u_{kj} = 0 \text{ (produto de duas linhas distintas)} \ .$$

2) $||Ux|| \ = \ ||x||, \ \forall x \in I\!R^n$.

3) A transformação ortogonal não muda os ângulos entre dois vetores. Portanto uma transformação ortogonal ou é uma rotação ou é uma reflexão.

4) Os auto-valores são: 1 ou -1.

5) O determinante é 1 ou -1.

Para finalizar essa seção daremos um teorema que nos permite ter uma idéia da localização dos auto-valores de uma matriz, seja ela simétrica ou não. Os auto-valores de matrizes não simétricas podem, é lógico, serem complexos, e nestes casos o teorema fornece a localização destes números no plano complexo. Existem situações onde não é necessário obter os auto-valores com muita precisão, isto é, existem ocasiões onde o que desejamos é saber se os auto-valores são positivos ou então se estão contidos no círculo unitário. O Teorema a seguir pode ser usado para responder a estas perguntas sem a necessidade de cálculos detalhados.

Teorema 1.10 - *Teoremas de Gerschgorin*

a) Primeiro Teorema de Gerschgorin - Os auto-valores de uma matriz $A = (a_{ij})$ estão na reunião dos círculos de centro a_{ii} e raio

$$r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \ , \quad i = 1, 2, \ldots, n \ ,$$

no plano complexo.

b) Segundo Teorema de Gerschgorin - Se a união de q desses círculos formam uma região conectada, isolada dos círculos restantes, então existe q auto-valores nessa região.

Prova: A prova deste teorema pode ser encontrada em [Wilkison, 1965].

Exemplo 1.24 - *Localizar, usando o teorema de Gerschgorin, os auto-valores de:*

$$A = \begin{pmatrix} 4 & -1 & 1 \\ 1 & 1 & 1 \\ -2 & 0 & -6 \end{pmatrix} , \quad B = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & -1 \\ 0 & -1 & 0 \end{pmatrix} .$$

Solução: Os círculos de Gerschgorin associados com a matriz A são dados por:

Círculo	Centro	Raio
C_1	$a_{11} = 4$	$r_1 = \|-1\| + \|1\| = 2$
C_2	$a_{22} = 1$	$r_2 = \|1\| + \|1\| = 2$
C_3	$a_{33} = -6$	$r_3 = \|-2\| + \|0\| = 2$

Assim para a matriz A, obtemos os círculos ilustrados na Figura 1.5:

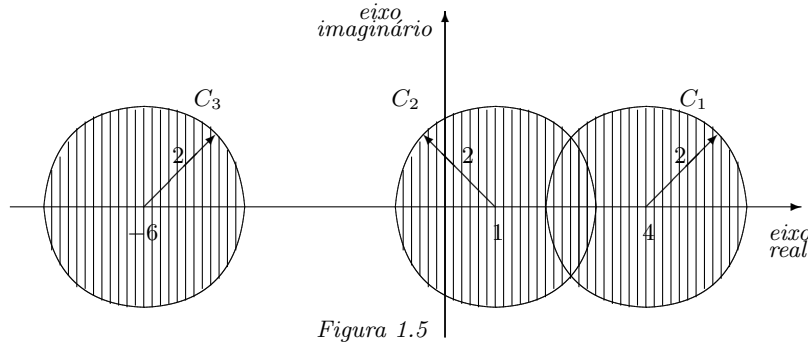



Figura 1.5

O primeiro teorema de Gerschgorin indica que os auto-valores de A estão inseridos nas regiões hachuradas da Figura 1.5. Além disso, desde que $C_1 \bigcup C_2$ não intercepta C_3, pelo segundo teorema de Gerschgorin, dois desses auto-valores estão em $C_1 \bigcup C_2$ e os restantes dos auto-valores em C_3.

Para a matriz B, temos que os círculos de Gerschgorin associados com essa matriz, são dados por:

Círculo	Centro	Raio
C_1	$b_{11} = 3$	$r_1 = \|1\| + \|0\| = 1$
C_2	$b_{22} = 2$	$r_2 = \|1\| + \|-1\| = 2$
C_3	$b_{33} = 0$	$r_3 = \|0\| + \|-1\| = 1$

os quais estão ilustrados na Figura 1.6.

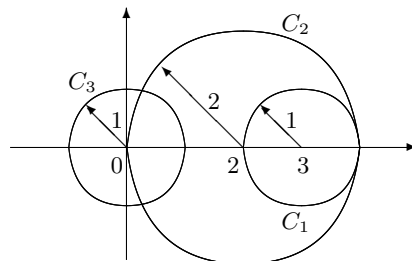



Figura 1.6

Podemos afirmar neste caso, usando os teoremas de Gerschgorin, que os auto-valores da matriz B estão no intervalo $[-1, 4]$, pois a matriz é real e simétrica.

Exercícios

1.27 - *Dada as seguintes matrizes:*

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \ , \ B = \begin{pmatrix} 1 & 2 & -1 \\ -1 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix} \ ,$$

calcule o polinômio característico, seus auto-valores e auto-vetores.

1.28 - *Seja* $A = \begin{pmatrix} 1 & 2 \\ 3 & 2 \end{pmatrix}$. *Calcule os auto-valores de* A, A^2, A^3.

1.29 - *Seja* $A = \begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}$. *Calcular* $P(A)$ *e* $Q(A)$, *sabendo que:* $P(t) = 2t^2 - 3t + 7$ *e* $Q(t) = t^2 - 5$.

1.6 Exercícios Complementares

1.30 - *Se* $x = (1,\ 2,\ 3,\ 4)^t$ *e* $y = (0,\ 3,\ -2,\ 1)^t$, *calcule:*

a) (x, y) *(usando definição usual de produto escalar),*

b) $\| x \|$ *e* $\| y \|$,

1.31 - *Mostre que num espaço euclidiano vale o Teorema de Pitágoras, isto é:*

$$x \perp y \implies \| x + y \|^2 = \| x \|^2 + \| y \|^2 \ .$$

1.32 - *Mostre que num espaço euclidiano, vale:*

$$| \, \| x \| - \| y \| \, | \leq \| x - y \| \ .$$

1.33 - *Sejam* $x = (x_1,\ x_2)^t$ *e* $y = (y_1, y_2)^t$ *vetores do* $I\!R^2$.

a) *Prove que:*

$$(x, y) = x_1 \, y_1 - 2 \, x_1 \, y_2 - 2 \, x_2 \, y_1 + 5 \, x_2 \, y_2 \ ,$$

define um produto escalar no $I\!R^2$. .

b) *Determine a norma de* $x = (1,\ 2)^t \in I\!R^2$, *em relação ao produto escalar do item a).*

1.34 - *Os vetores* $\{(1,\ 1,\ 0)^t, (0,\ 1,\ 1)^t, (1,\ 0,\ 1)^t\}$ *constituem uma base não ortonormal do* $I\!R^3$. *Construir a partir desses vetores, uma base ortonormal para o* $I\!R^3$, *usando o processo de Gram-Schmidt.*

1.35 - *Obter, no intervalo* $[0, 1]$, *uma sequência ortonormal de polinômios, relativamente ao produto escalar.:*

$$(f, g) = \int_0^1 f(x) \ g(x) \ dx \ .$$

1.36 - *Considere o espaço dos polinômios de grau* ≤ 2 *com o produto escalar:*

$$(P_i, P_j) = \int_0^1 P_i(t) \ P_j(t) \ dt \ .$$

Dada nesse espaço a base $\{3,\ t - 3,\ t^2 - t\}$, *obtenha a partir dela uma base ortogonal, usando o processo de Gram-Schmidt.*

1.37 - *Sejam* $e_1,\ e_2,\ e_3$ *a base canônica do* $I\!R^3$, *e seja* $v = (1,\ 1,\ 2)^t$. *Determinar a projeção ortogonal de* v *sobre o plano* $\{e_1,\ e_2\}$.

1.38 - *Seja* $E = C[1,2]$, *com* $(f,g) = \int_1^2 f(x)g(x)dx$. *Seja* $K_1(x)$ *o sub-espaço dos polinômios de grau* ≤ 1. *O conjunto* $\{1,\ x\}$ *constitui uma base de* $K_1(x)$. *Determinar a projeção ortogonal de* $f(x) = e^x$ *sobre* $k_1(x)$.

1.39 - *Resolva o exercício 1.24, usando para o sub-espaço a base ortogonal obtida no exercício 1.36. Compare os resultados.*

1.40 - *Para cada uma das matrizes:*

$$A = \begin{pmatrix} -2 & 5 \\ 1 & -3 \end{pmatrix} \ , \quad A = \begin{pmatrix} 1 & 4 & 3 \\ 0 & 3 & 1 \\ 0 & 2 & -1 \end{pmatrix} \ ,$$

encontre um polinômio que tenha a matriz como raiz.

1.41 - *Seja* A *uma matriz quadrada de ordem* n *e sejam* $\lambda_1, \lambda_2, \cdots, \lambda_n$ *seus auto-valores. Quais são os auto-valores de* $A - qI$ *onde* q *é uma constante e* I *é a matriz identidade?*

1.42 - *Mostre que se* v *é auto-vetor de* A *e de* B *então* v *é auto-vetor de* $\alpha A + \beta B$, *onde* α, β *são escalares quaisquer.*

1.43 - *Mostre que uma matriz* A *e sua transposta* A^t *possuem o mesmo polinômio característico.*

1.44 - *Usando o Teorema 1.10, localizar os auto-valores das seguintes matrizes:*

$$A \ = \ \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \ , \quad B \ = \ \begin{pmatrix} 4 & 0 & 1 \\ -2 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} \ .$$

Capítulo 2

Análise de Arredondamento em Ponto Flutuante

2.1 Introdução

Neste capítulo, chamamos atenção para o fato de que o conjunto dos números representáveis em qualquer máquina é finito, e portanto *discreto*, ou seja não é possível representar em uma máquina todos os números de um dado intervalo $[a, b]$. A implicação imediata desse fato é que o resultado de uma simples operação aritmética ou o cálculo de uma função, realizadas com esses números, podem conter erros. A menos que medidas apropriadas sejam tomadas, essas imprecisões causadas, por exemplo, por simplificação no modelo matemático (algumas vezes necessárias para se obter um modelo matemático solúvel); erro de truncamento (troca de uma série infinita por uma finita); erro de arredondamento (devido a própria estrutura da máquina); erro nos dados (dados imprecisos obtidos de experimentos, ou arredondados na entrada); etc, podem diminuir e algumas vezes destruir, a precisão dos resultados, mesmo em precisão dupla.

Assim, nosso objetivo aqui será o de alertar o leitor para os problemas que possam surgir durante a resolução de um problema, bem como dar subsídios para evitá-los e para uma melhor interpretação dos resultados obtidos.

2.2 Sistema de Números Discreto no Computador

Inicialmente, descreveremos como os números são representados num computador.

Representação de um Número Inteiro

Em princípio, a representação de um **número inteiro** no computador não apresenta qualquer dificuldade. Qualquer computador trabalha internamente com uma base fixa β, onde β é um inteiro ≥ 2; e é escolhido como uma potência de 2.

Assim dado um número inteiro $n \neq 0$, ele possui uma única representação,

$$n \; = \; \pm(n_{-k}n_{-k+1}\ldots n_{-1}n_0) = \pm(n_0\beta^0 + n_{-1}\beta^1 + \ldots n_{-k}\beta^k),$$

onde os $n_i, \; i = 0, -1, \ldots, -k$ são inteiros satisfazendo $0 \leq n_i < \beta$ e $n_{-k} \neq 0$.

Por exemplo, na base $\beta = 10$, o número 1997 é representado por:

$$1997 \;=\; 7 \times 10^0 + 9 \times 10^1 + 9 \times 10^2 + 1 \times 10^3 \;,$$

e é armazenado como $n_{-3}n_{-2}n_{-1}n_0$.

Representação de um número real

A representação de um **número real** no computador pode ser feita de duas maneiras:

a) Representação em ponto fixo

Este foi o sistema usado, no passado, por muitos computadores. Assim, dado um número real, $x \neq 0$, ele será representado em **ponto fixo** por:

$$x \;=\; \pm \sum_{i=k}^{n} x_i \; \beta^{-i} \;,$$

onde k e n são inteiros satisfazendo $k < n$ e usualmente $k \leq 0$ e $n > 0$ e os x_i são inteiros satisfazendo $0 \leq x_i < \beta$.

Por exemplo, na base $\beta = 10$, o número 1997.16 é representado por:

$$\begin{aligned} 1997.16 &= \sum_{i=-3}^{2} x_i \beta^{-i} \\ &= 1 \times 10^3 + 9 \times 10^2 + 9 \times 10^1 + 7 \times 10^0 + 1 \times 10^{-1} + 6 \times 10^{-2} \\ &= 1 \times 1000 + 9 \times 100 + 9 \times 10 + 7 \times 1 + 1 \times 0.1 + 6 \times 0.01 \;, \end{aligned}$$

e é armazenado como $x_{-3}x_{-2}x_{-1}x_0.x_1x_2$.

b) Representação em Ponto Flutuante

Esta representação, que é mais flexível que a representação em ponto fixo, é universalmente utilizada nos dias atuais. Dado um número real, $x \neq 0$, este será representado em **ponto flutuante** por:

$$x \;=\; \pm \; d \times \beta^e \;,$$

onde β é a base do sistema de numeração, d é a mantissa e e é o expoente. A mantissa é um número em ponto fixo, isto é:

$$d \;=\; \sum_{i=k}^{n} d_i \; \beta^{-i} \;,$$

onde, frequentemente, nos grandes computadores, $k = 1$, tal que se $x \neq 0$, então $d_1 \neq 0$; $0 \leq d_i < \beta$, $i = 1, 2, \ldots t$, com t a quantidade de dígitos significativos ou precisão do sistema , $\beta^{-1} \leq d < 1$ e $-m \leq e \leq M$.

Observações:

a) $d_1 \neq 0$ caracteriza o sistema de números em ponto flutuante **normalizado**.

b) o número **zero** pertence a qualquer sistema e é representado com mantissa igual a zero e $e \;=\; -m$.

Exemplo 2.1 - *Escrever os números:*

$$x_1 = 0.35; \quad x_2 = -5.172; \; x_3 = 0.0123; \; x_4 = 5391.3 \text{ e } x_5 = 0.0003 \; ,$$

onde todos estão na base $\beta = 10$*, em ponto flutuante na forma normalizada.*

Solução: Temos então:

$$\begin{aligned}
0.35 &= (3 \times 10^{-1} + 5 \times 10^{-2}) \times 10^0 = 0.35 \times 10^0 \; , \\
-5.172 &= -(5 \times 10^{-1} + 1 \times 10^{-2} + 7 \times 10^{-3} + 2 \times 10^{-4}) \times 10^1 \\
&= -0.5712 \times 10^1 \; , \\
0.0123 &= (1 \times 10^{-1} + 2 \times 10^{-2} + 3 \times 10^{-3}) \times 10^{-1} = 0.123 \times 10^{-1} \; , \\
5391.3 &= (5 \times 10^{-1} + 3 \times 10^{-2} + 9 \times 10^{-3} + 1 \times 10^{-4} + 3 \times 10^{-5}) \times 10^4 \\
&= 0.53913 \times 10^4 \; , \\
0.0003 &= (3 \times 10^{-1}) \times 10^{-3} = 0.3 \times 10^{-3} \; .
\end{aligned}$$

Agora, para representarmos um sistema de números em ponto flutuante normalizado, na base β, com t dígitos significativos e com limites do expoente m e M , usaremos a notação: $\mathbf{F}(\beta, \mathbf{t}, \mathbf{m}, \mathbf{M})$.

Assim um número em $F(\beta, t, m, M)$ será representado por:

$$\pm \; 0.d_1 d_2 \ldots d_t \times \beta^e \; ,$$

onde $d_1 \neq 0$ e $-m \leq e \leq M$.

Exemplo 2.2 - *Considere o sistema* $F(10, 3, 2, 2)$*. Represente nesse sistema os números do exemplo anterior.*

Solução: Temos então que nesse sistema um número será representado por $\pm \; 0.d_1 d_2 d_3 \times 10^e$, onde $-2 \leq e \leq 2$. Assim:

$$\begin{aligned}
0.35 &= 0.350 \times 10^0 \; , \\
-5.172 &= -0.517 \times 10^1 \; , \\
0.0123 &= 0.123 \times 10^{-1} \; ,
\end{aligned}$$

Observe que os números 5391.3 e 0.0003 não podem ser representados no sistema. De fato, o número $5391.3 = 0.539 \times 10^4$ e portanto o expoente é maior que 2, causando *overflow*, por outro lado $0.0003 = 0.300 \times 10^{-3}$ e assim o expoente é menor que -2 causando *underflow*.

Podemos então definir formalmente dígitos significativos de um número.

Definição 2.1 - *Seja* β *a base do sistema de números em ponto flutuante.* **Dígitos significativos** *de um número* x*, são todos os algarismos de 0 a* $\beta - 1$*, desde que* x *esteja representado na forma normalizada.*

Para exemplificar as limitações da máquina, consideremos agora o seguinte exemplo.

Exemplo 2.3 - *Seja* $f(x)$ *uma função contínua real definida no intervalo* $[a, b]$*,* $a < b$ *e sejam* $f(a) < 0$ *e* $f(b) > 0$*. Então de acordo com o teorema do valor intermediário, existe* x*,* $a < x < b$ *tal que* $f(x) = 0$*. Seja* $f(x) = x^3 - 3$*. Determinar* x *tal que* $f(x) = 0$*.*

Solução: Para a função dada , consideremos $t = 10$ e $\beta = 10$. Obtemos então:

$$\begin{aligned}
&f(0.1442249570 \times 10^1) = -0.2 \times 10^{-8} \; ; \\
&f(0.1442249571 \times 10^1) = 0.4 \times 10^{-8} \; .
\end{aligned}$$

Observe que entre 0.1442249570×10^1 e 0.1442249571×10^1 não existe nenhum número que possa ser representado no sistema dado e que a função f muda de sinal nos extremos desse intervalo. Assim, esta máquina não contém o número x tal que $f(x) = 0$ e portanto a equação dada não possui solução.

Exercícios

2.1 - *Considere o sistema* $F(10,4,4,4)$. *Represente neste sistema os números:* $x_1 = 4321.24$, $x_2 = -0.0013523$, $x_3 = 125.64$, $x_4 = 57481.23$ *e* $x_5 = 0.00034$.

2.2 - *Represente no sistema* $F(10,3,1,3)$ *os números do exercício 2.1.*

Mudança de Base

Como já dissemos anteriormente a maioria dos computadores trabalham na base β onde β é um inteiro ≥ 2; e é normalmente escolhido como uma potência de 2. Assim um mesmo número pode ser representado em mais do que uma base. Além disso sabemos que, através de uma **mudança de base**, é sempre possível determinar a representação em uma nova base. Veremos então, através de exemplos, como se faz mudança de base.

Exemplo 2.4 - *Mudar a representação dos números:*

i) 1101 *da base 2, para a base 10,*

ii) 0.110 *da base 2, para a base 10,*

iii) 13 *da base 10, para a base 2,*

iv) 0.75 *da base 10, para a base 2,*

v) 3.8 *da base 10, para a base 2.*

Solução: Para cada número daremos qual o procedimento a ser seguido. Assim:

i) 1101 que está na base 2, para a base 10.

Neste caso o procedimento é multiplicar cada algarismo do número na base 2 por potências crescente de 2, da direita para a esquerda e somar todas as parcelas. Assim:

$$1101 \;=\; 1 \times 2^0 + 0 \times 2^1 + 1 \times 2^2 + 1 \times 2^3 = 1 + 0 + 4 + 8 = 13\ .$$

Logo, $(1101)_2 \;=\; (13)_{10}$.

ii) 0.110 que está na base 2, para a base 10.

Neste caso o procedimento é multiplicar cada algarismo do número na base 2, após o ponto, por potências decrescente de 2, da esquerda para a direita e somar todas as parcelas. Assim:

$$0.110 \;=\; 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} = \frac{1}{2} + \frac{1}{4} + 0 = 0.75\ .$$

Logo, $(0.110)_2 \;=\; (0.75)_{10}$.

iii) 13 que está na base 10, para a base 2.

Neste caso o procedimento é dividir o número por 2. A seguir continuar dividindo o quociente por 2 até que o último quociente seja igual a 1. O número na base 2 será então obtido tomando-se o último quociente e todos os restos das divisões anteriores. Assim:

$$\begin{array}{llll} 13 & \underline{|\ \ 2\ \ } & & \\ \mathbf{1} & 6 & \underline{|\ \ 2\ \ } & \\ & \mathbf{0} & 3 & \underline{|\ \ 2\ \ } \\ & & \mathbf{1} & \mathbf{1} \end{array}$$

Logo, $(13)_{10} = (1101)_2$.

iv) 0.75 que está na base 10, para a base 2.

Neste caso o procedimento é multiplicar a parte decimal por 2. A seguir continuar multiplicando a parte decimal do resultado obtido, por 2. O número na base 2 será então obtido tomando-se a parte inteira do resultado de cada multiplicação. Assim:

$$\begin{aligned} 0.75 \times 2 &= \mathbf{1}.50 \\ 0.50 \times 2 &= \mathbf{1}.00 \\ 0.00 \times 2 &= \mathbf{0}.00 \end{aligned}$$

Logo, $(0.75)_{10} = (0.110)_2$.

v) 3.8 que está na base 10, para a base 2.

O procedimento neste caso é transformar a parte inteira seguindo o item **iii)** o que nos fornece $(3)_{10} = (11)_2$ e a parte decimal seguindo o item **iv)**. Assim, obtemos:

$$\begin{aligned} 0.8 \times 2 &= \mathbf{1}.6 \\ 0.6 \times 2 &= \mathbf{1}.2 \\ 0.2 \times 2 &= \mathbf{0}.4 \\ 0.4 \times 2 &= \mathbf{0}.8 \\ 0.8 \times 2 &= \ldots \end{aligned}$$

Logo, $(3.8)_{10} = (11.11001100\ldots)_2$. Portanto o número $(3.8)_{10}$ não tem representação exata na base 2. Esse exemplo ilustra também o caso de erro de arredondamento nos dados.

No exemplo 2.4, mudamos a representação de números na base 10 para a base 2 e vice-versa. O mesmo procedimento pode ser utilizado para mudar da base 10 para uma outra base qualquer e vice-versa. A pergunta que surge naturalmente é: qual o procedimento para representar um número que está numa dada base β_1 em uma outra base β_2, onde $\beta_1 \neq \beta_2 \neq 10$? Nesse caso devemos seguir o seguinte procedimento: inicialmente representamos o número que está na base β_1 na base 10 e a seguir o número obtido na base 10, na base β_2.

Exemplo 2.5 - *Dado o número* **12.20** *que está na base 4, representá-lo na base 3.*

Solução: Assim, usando os procedimentos dados no exemplo 2.4, obtemos:

$$\begin{aligned} 12 &= 2 \times 4^0 + 1 \times 4^1 = 6. \\ 0.20 &= 2 \times 4^{-1} + 0 \times 4^{-2} = \frac{2}{4} = 0.5\ . \end{aligned}$$

Portanto: $(12.20)_4 = (6.5)_{10}$. Agora:

$$\begin{array}{l|l} 6 & 3 \\ \hline \mathbf{0} & 2 \end{array}$$

$$\begin{aligned} 0.5 \times 3 &= \mathbf{1}.5 \\ 0.5 \times 3 &= \mathbf{1}.5 \\ &\vdots \end{aligned}$$

Portanto: $(6.5)_{10} = (20.11\ldots)_3$. Logo $(12.20)_4 = (20.111\ldots)_3$. Observe que o número dado na base 4, tem representação exata na base 10, mas não na base 3.

Exercícios

2.3 - *Considere os seguintes números:* $x_1 = 34$, $x_2 = 0.125$ *e* $x_3 = 33.023$ *que estão na base 10. Escreva-os na base 2.*

2.4 - *Considere os seguintes números:* $x_1 = 110111$, $x_2 = 0.01011$ *e* $x_3 = 11.0101$ *que estão na base 2. Escreva-os na base 10.*

2.5 - *Considere os seguintes números:* $x_1 = 33$, $x_2 = 0.132$ *e* $x_3 = 32.013$ *que estão na base 4. Escreva-os na base 5.*

2.3 Representação de Números no Sistema $F(\beta, t, m, M)$

Sabemos que os números reais podem ser representados por uma reta contínua. Entretanto, em ponto flutuante podemos representar apenas pontos discretos na reta real. Para ilustrar este fato consideremos o seguinte exemplo.

Exemplo 2.6 - *Quantos e quais números podem ser representados no sistema* $F(2, 3, 1, 2)$*?*

Solução: Temos que $\beta = 2$ então os dígitos podem ser 0 ou 1; $m = 1$ e $M = 2$ então $-1 \leq e \leq 2$ e $t = 3$. Assim, os números são da forma :

$$\pm\ 0.d_1 d_2 d_3 \times \beta^e \ .$$

Logo temos: duas possiblidades para o sinal, uma possiblidade para d_1, duas para d_2 , duas para d_3 e quatro para as formas de β^e. Fazendo o produto $2 \times 1 \times 2 \times 2 \times 4$ obtemos 32. Assim neste sistema podemos representar 33 números visto que o zero faz parte de qualquer sistema.

Para responder quais são os números, notemos que as formas da mantissa são : 0.100, 0.101, 0.110 e 0.111 e as formas de β^e são: 2^{-1}, 2^0, 2^1, 2^2. Assim, obtemos os seguintes números:

$$0.100 \times \left\{ \begin{array}{rcl} 2^{-1} &=& (0.25)_{10} \\ 2^0 &=& (0.5)_{10} \\ 2^1 &=& (1.0)_{10} \\ 2^2 &=& (2.0)_{10} \ , \end{array} \right.$$

desde que $(0.100)_2 \ = \ (0.5)_{10}$;

$$0.101 \times \left\{ \begin{array}{rcl} 2^{-1} &=& (0.3125)_{10} \\ 2^0 &=& (0.625)_{10} \\ 2^1 &=& (1.25)_{10} \\ 2^2 &=& (2.5)_{10} \ , \end{array} \right.$$

desde que $(0.101)_2 \;=\; (0.625)_{10}$;

$$0.110 \times \begin{cases} 2^{-1} & = & (0.375)_{10} \\ 2^0 & = & (0.75)_{10} \\ 2^1 & = & (1.5)_{10} \\ 2^2 & = & (3.0)_{10}\ , \end{cases}$$

desde que $(0.110)_2 \;=\; (0.75)_{10}$;

$$0.111 \times \begin{cases} 2^{-1} & = & (0.4375)_{10} \\ 2^0 & = & (0.875)_{10} \\ 2^1 & = & (1.75)_{10} \\ 2^2 & = & (3.5)_{10}\ , \end{cases}$$

desde que $(0.111)_2 \;=\; (0.875)_{10}$.

Exemplo 2.7 - *Considerando o mesmo sistema do exemplo 2.6, represente os números:* $x_1 = 0.38$, $x_2 = 5.3$ *e* $x_3 = 0.15$ *dados na base 10.*

Solução: Fazendo os cálculos obtemos que: $(0.38)_{10} = 0.110 \times 2^{-1}$, $(5.3)_{10} = 0.101 \times 2^3$ e $(0.15)_{10} = 0.100 \times 2^{-2}$. Assim apenas o primeiro número pode ser representado no sistema, pois para o segundo teremos *overflow* e para o terceiro *underflow*. Observe que o número $(0.38)_{10}$ tem no sistema dado, a mesma representação que o número $(0.375)_{10}$.

Exercícios

2.6 - *Considere o sistema* $F(3,3,2,1)$.

a) *Quantos e quais números podemos representar neste sistema?*

b) *Represente no sistema os números:* $x_1 = (0.40)_{10}$, $x_2 = (2.8)_{10}$.

2.7 - *Considere o sistema* $F(2,5,3,1)$.

a) *Quantos números podemos representar neste sistema?*

b) *Qual o maior número na base 10 que podemos representar neste sistema (sem fazer arredondamento)?*

Todas as operações num computador são *arredondadas*. Para ilustrar este fato, consideremos o seguinte exemplo.

Exemplo 2.8 - *Calcular o quociente entre 15 e 7.*

Solução: Temos três representações alternativas:

$$x_1 \;=\; \frac{15}{7}, \quad x_2 \;=\; 2\frac{1}{7}, \quad x_3 \;=\; 2.142857.$$

Note que x_1 e x_2 são representações exatas e x_3 é uma aproximação do quociente.

Suponha agora que só dispomos de 4 dígitos para representar o quociente $\frac{15}{7}$. Daí, $\frac{15}{7} \;=\; 2.142$. Mas não seria melhor aproximarmos $\frac{15}{7}$ por 2.143? A resposta é sim e isso significa que o número foi arredondado. Mas o que significa arredondar um número?

Arredondamento em Ponto Flutuante

Definição 2.2 - *Arredondar um número x, por outro com um número menor de dígitos significativos, consiste em encontrar um número $\bar{x}$, pertecente ao sistema de numeração, tal que $|\bar{x} - x|$ seja o menor possível.*

Assim para o exemplo dado $|2.142 - x_3| = 0.000857$ e $|2.143 - x_3| = 0.000143$. Logo 2.143 representa a melhor aproximação para $\frac{15}{7}$, usando 4 dígitos significativos.

Daremos a seguir a regra de como arredondar um número.

Dado x, seja $\bar{x}$ sua representação em $F(\beta, t, m, M)$ adotando arredondamento. Se $x = 0$ então $\bar{x} = 0$. Se $x \neq 0$, então escolhemos s e e tais que:

$$|x| \;=\; s \times \beta^e \quad \text{onde} \quad \beta^{-1}(1 - \frac{1}{2}\beta^{-t}) \leq s < 1 - \frac{1}{2}\beta^{-t}. \tag{2.1}$$

Se e está fora do intervalo $[-m, M]$ não temos condições de representar o número no sistema. Se $e \in [-m, M]$ então calculamos:

$$s + \frac{1}{2}\beta^{-t} \;=\; 0.d_1 d_2 \ldots d_t d_{t+1} \ldots$$

e truncamos em t dígitos. Assim o número arredondado será:

$$\bar{x} \;=\; (sinal x)(0.d_1 d_2 \ldots d_t) \times \beta^e.$$

Exemplo 2.9 - *Considere o sistema $F(10, 3, 5, 5)$. Represente neste sistema os números: $x_1 = 1234.56$, $x_2 = -0.00054962$, $x_3 = 0.9995$, $x_4 = 123456.7$ e $x_5 = -0.0000001$.*

Solução: Primeiramente, analisemos quais os valores permitidos para s. Desde que $\beta = 10$ e $t = 3$, usando (2.1), segue que:

$$10^{-1}(1 - \frac{1}{2}10^{-3}) \leq s < 1 - \frac{1}{2}10^{-3} \ ,$$

e fazendo os cálculos obtemos que:

$$0.09995 \leq s < 0.9995 \ .$$

Podemos agora tentar representar os números no sistema dado. Assim:

i) Para $x_1 = 1234.56$, obtemos:

$$|x_1| \;=\; 0.123456 \times 10^4,$$

$$s + \frac{1}{2}10^{-3} \;=\; 0.123456 + 0.0005 \;=\; 0.123956,$$

$$\bar{x}_1 \;=\; 0.123 \times 10^4 \ ;$$

ii) para $x_2 = -0.00054962$, obtemos:

$$|x_2| \;=\; 0.54962 \times 10^{-3},$$

$$s + \frac{1}{2}10^{-3} \;=\; 0.54962 + 0.0005 = 0.55012,$$

$$\bar{x}_2 \;=\; -0.550 \times 10^{-3} \ ;$$

iii) para $x_3 = 0.9995$, observe que não podemos considerar $|x_3| = 0.9995 \times 10^0$, pois neste caso s não pertence ao seu intervalo, e o número arredondado não estaria escrito na forma dos elementos do sistema. Assim neste caso consideramos:

$$\begin{aligned} |x_3| &= 0.09995 \times 10^1, \\ s + \frac{1}{2}10^{-3} &= 0.09995 + 0.0005 = 0.10045, \\ \bar{x}_3 &= 0.100 \times 10^1 \ ; \end{aligned}$$

iv) para $x_4 = 123456.7$, obtemos:

$$|x_4| = 0.1234567 \times 10^6,$$

v) para $x_5 = -0.0000001$, obtemos:

$$|x_5| = 0.1 \times 10^{-6} \ .$$

Observe que tanto em **iv)** como em **v)** não podemos representar o número no sistema dado pois em **iv)** teremos *overflow* e em **v)** *underflow*.

Assim, em linhas gerais, para arredondar um número, na base 10, devemos apenas observar o primeiro dígito a ser descartado. Se este dígito é menor que 5 deixamos os dígitos inalterados e se é maior ou igual a 5 devemos somar 1 ao último dígito remanescente.

Exercício

2.8 - *Considere o sistema $F(10, 4, 4, 4)$.*

a) *Qual o intervalo para s neste caso?*

b) *Represente os números do exemplo 2.9 nesse sistema.*

2.4 Operações Aritméticas em Ponto Flutuante

Considere uma máquina qualquer e uma série de operações aritméticas. Pelo fato do arredondamento ser feito após cada operação temos, ao contrário do que é válido para números reais, que as operações aritméticas (adição, subtração, divisão e multiplicação) não são nem associativas e nem distributivas. Ilustraremos esse fato através de exemplos.

Nos exemplos desta seção considere o sistema com base $\beta = 10$, e 3 dígitos significativos.

Exemplo 2.10 - *Efetue as operações indicadas:*

$$\begin{array}{ll} i) & (11.4 + 3.18) + 5.05 \ e \ 11.4 + (3.18 + 5.05) \ , \\ ii) & \dfrac{3.18 \times 11.4}{5.05} \ e \ \left(\dfrac{3.18}{5.05}\right) \times 11.4 \ , \\ iii) & 3.18 \times (5.05 + 11.4) \ e \ 3.18 \times 5.05 + 3.18 \times 11.4 \ . \end{array}$$

Solução: Para cada item, fazendo o arredondamento após cada uma das operações efetuada, segue que:

$$\begin{array}{rl} \text{i)} & (11.4 + 3.18) + 5.05 \; = \; 14.6 + 5.05 \; = \; 19.7 \; , \\ \textit{enquanto} & \\ & 11.4 + (3.18 + 5.05) \; = \; 11.4 + 8.23 \; = \; 19.6 \; . \\ \text{ii)} & \dfrac{3.18 \times 11.4}{5.05} \; = \; \dfrac{36.3}{5.05} \; = \; 7.19 \; , \\ \textit{enquanto} & \\ & \left(\dfrac{3.18}{5.05}\right) \times 11.4 \; = \; 0.630 \times 11.4 \; = \; 7.18 \; . \\ \text{iii)} & 3.18 \times (5.05 + 11.4) \; = \; 3.18 \times 16.5 \; = \; 52.3 \; , \\ \textit{enquanto} & \\ & 3.18 \times 5.05 + 3.18 \times 11.4 \; = \; 16.1 + 36.3 \; = \; 52.4 \; . \end{array}$$

Exemplo 2.11 - *Somar* $\frac{1}{3}$ *dez vezes consecutivas usando arredondamento.*

Solução: Temos então:

$$\underbrace{0.333 + 0.333 + \ldots + 0.333}_{10 \text{ vezes}} \; = \; 3.31 \; .$$

Entretanto podemos obter um resultado melhor se multiplicarmos 0.333 por 10, obtendo assim 3.33.

Exemplo 2.12 - *Avaliar o polinômio*

$$P(x) \; = \; x^3 \; - \; 6\, x^2 \; + \; 4\, x \; - \; 0.1 \; ,$$

no ponto 5.24 *e comparar com o resultado exato.*

Solução: Para calcular o valor *exato* consideremos todos os dígitos de uma máquina, sem usar arredondamento a cada operação. Assim:

$$P(5.24) \; = \; 143.8777824 \; - \; 164.7456 \; + \; 20.96 \; - \; 0.1 \; = \; -0.00776 \quad (\text{ valor exato}).$$

Agora, usando arredondamento a cada operação efetuada, obtemos:

$$\begin{aligned} P(5.24) \; &= \; 5.24 \times 27.5 \; - \; 6 \times 27.5 \; + \; 4 \times 5.24 \; - \; 0.1 \\ &= \; 144. \; - \; 165. \; + \; 21.0 \; - \; 0.1 \\ &= \; -0.10 \quad (\text{somando da esquerda para a direita}) \\ &= \; 0.00 \quad (\text{somando da direita para a esquerda}). \end{aligned}$$

Entretanto, observe que $P(x)$ pode ser escrito como:

$$P(x) \; = \; x\,(x\,(x \; - \; 6) \; + \; 4) \; - \; 0.1 \; .$$

Assim:

$$\begin{aligned} P(5.24) \; &= \; 5.24\,(5.24\,(5.24 \; - \; 6) \; + \; 4) \; - \; 0.1 \\ &= \; 5.24\,(-3.98 \; + \; 4) \; - \; 0.1 \\ &= \; 5.24\,(0.02) \; - \; 0.1 \\ &= \; 0.105 \; - \; 0.1 \\ &= \; 0.005 \quad (\text{sinal errado}). \end{aligned}$$

Observando os três últimos exemplos, vemos que erros consideráveis podem ocorrer durante e execução de um algoritmo. Isso se deve ao fato de que existem limitações da máquina e também que os erros de arredondamento são introduzidos a cada operação efetuada. Em consequência, podemos obter resultados diferentes mesmo utilizando métodos numéricos matematicamente equivalentes.

Assim, devemos ser capazes de conseguir desenvolver um algoritmo tal que os efeitos da aritmética discreta do computador permaneça inofensivo quando um grande número de operações são executadas.

Exercícios

2.9 - *Considere o sistema $F(10, 3, 5, 5)$. Efetue as operações indicadas:*

$$i) \quad (1.386 - 0.987) + 7.6485 \quad e \quad 1.386 - (0.987 - 7.6485) \ ,$$

$$ii) \quad \frac{1.338 - 2.038}{4.577} \quad e \quad \left(\frac{1.338}{4.577}\right) - \left(\frac{2.038}{4.577}\right) \ ,$$

2.10 - *Seja*

$$x = \frac{17.678}{3.471} + \frac{(9.617)^2}{3.716 \times 1.85}$$

a) *Calcule x com todos os algarismos da sua calculadora, sem efetuar arredondamento.*

b) *Calcule x considerando o sistema $F(10, 3, 4, 3)$. Faça arredondamento a cada operação efetuada.*

2.11 - *Seja $P(x) = 2.3\ x^3 - 0.6\ x^2 + 1.8\ x - 2.2$. Deseja-se obter o valor de $P(x)$ para $x = 1.61$.*

a) *Calcule $P(1.61)$ com todos os algarismos da sua calculadora, sem efetuar arredondamento.*

b) *Calcule $P(1.61)$ considerando o sistema $F(10, 3, 4, 3)$. Faça arredondamento a cada operação efetuada.*

2.12 - *Seja:*

$$S \ = \ \sum_{i=1}^{n} = \frac{n(n+1)}{2} \ .$$

Calcule S, considerando $n = 1000$ e efetuando a soma dos termos em:

a) *ordem crescente,*

b) *ordem decrescente.*

2.5 Efeitos Numéricos

Além dos problemas dos erros causados pelas operações aritméticas, das fontes de erros citadas no início deste capítulo, existem certos efeitos numéricos que contribuem para que o resultado obtido não tenha crédito. Alguns dos mais frequentes são:

- Cancelamento
- Propagação do Erro
- Instabilidade Numérica
- Mal Condicionamento

2.5.1 Cancelamento

O **cancelamento** ocorre na subtração de dois números quase iguais. Vamos supor que estamos operando com aritmética de ponto flutuante. Sejam x e y dois números com expoente e. Quando formamos a diferença $x - y$ ela também terá o expoente e. Se normalizarmos o número obtido, veremos que devemos mover os dígitos para a esquerda de tal forma que o primeiro seja diferente de zero. Assim, uma quantidade de dígitos iguais a zero aparecem no final da mantissa do número normalizado. Estes zeros não possuem significado algum.Veremos este fato através de exemplos, onde iremos considerar que estamos trabalhando com o sistema $F(10, 10, 10, 10)$.

Exemplo 2.13 - *Calcular:*

$$\sqrt{9876} - \sqrt{9875} \ .$$

Solução: Temos que:

$$\sqrt{9876} \ = \ 0.9937806599 \times 10^2 \quad \text{e} \quad \sqrt{9875} \ = \ 0.9937303457 \times 10^2 \ .$$

Portanto:

$$\sqrt{9876} - \sqrt{9875} \ = \ 0.0000503142 \times 10^2 \ .$$

A normalização muda este resultado para: $0.5031420000 \times 10^{-4}$. Assim os quatro zeros no final da mantissa não têm significado e assim perdemos 4 casas decimais. A pergunta que surge naturalmente é: podemos obter um resultado mais preciso? Neste caso a resposta é sim. Basta consideramos a identidade:

$$\sqrt{x} - \sqrt{y} \ = \ \frac{x - y}{\sqrt{x} + \sqrt{y}} \ ,$$

e assim, no nosso caso, obtemos:

$$\sqrt{9876} - \sqrt{9875} \ = \ \frac{1}{\sqrt{9876} + \sqrt{9875}} \ = \ 0.5031418679 \times 10^{-4} \ .$$

que é um resultado com todos os dígitos corretos.

Exemplo 2.14 - *Resolver a equação:*

$$x^2 - 1634 \ x + 2 = 0 \ .$$

Solução: Temos:

$$\begin{aligned} x \ &= \ \frac{1634 \pm \sqrt{(1634)^2 - 4(2)}}{2} \\ &= \ 817 \pm \sqrt{667487} \ . \end{aligned}$$

Assim:

$$\begin{aligned} x_1 \ &= \ 817 + 816.9987760 \ = \ 0.1633998776 \times 10^3 \ , \\ x_2 \ &= \ 817 - 816.9987760 \ = \ 0.1224000000 \times 10^{-2} \ . \end{aligned}$$

Os seis zeros da mantissa de x_2 são resultado do cancelamento e portanto não têm significado algum. Uma pergunta que surge naturalmente é: podemos obter um resultado mais preciso? Neste caso a resposta é sim. Basta lembrar que o produto das raízes é igual ao termo independente da equação, ou seja:

$$x_1 \times x_2 \ = \ 2 \ \rightarrow x_2 \ = \ \frac{2}{x_1} \ .$$

Logo: $x_2 = 0.1223991125 \times 10^{-2}$, onde agora todos os dígitos estão corretos.

Nos exemplos dados foi razoavelmente fácil resolver o problema do cancelamento. Entretanto, cabe salientar, que nem sempre existe uma maneira trivial de resolver problemas ocasionados pelo cancelamento.

2.5.2 Propagação do erro

O cancelamento não ocorre somente quando dois números quase iguais são subtraidos diretamente um do outro. Ele também ocorre no cálculo de uma soma quando uma soma parcial é muito grande quando comparada com o resultado final. Para exemplificar, consideremos que:

$$s_k = \sum_{k=1}^{n} a_k \ ,$$

seja a soma a ser calculada, onde os a_k podem ser positivos ou negativos. Vamos supor que o cálculo seja feito através de uma sequência de somas parciais da seguinte forma:

$$s_1 = a_1 \ , \quad s_k = s_{k-1} + a_k \ , \quad k = 2, 3, \ldots, n \ ,$$

tal que $s = s_n$.

Se a soma é calculada em aritmética de ponto fixo, então cada a_k está afetado de algum erro, os quais são limitados por algum ϵ para todo k. Se nenhum *overflow* ocorre, o erro na soma final s será de no máximo $n\epsilon$. Agora devido ao fato de nem todos os a_k terem o mesmo sinal então o erro será menor do que $n\epsilon$.

Mas se a soma é calculada em aritmética de ponto flutuante um novo fenômeno pode ocorrer. Vamos supor que uma das somas intermediárias s_k é consideravelmente grande em relação à soma final s, no sentido que o expoente de s_k excede o expoente de s em, digamos, p unidades. É claro que isso só pode ocorrer se nem todos os a_k possuem o mesmo sinal. Se simularmos tal soma em aritmética de ponto fixo (usando para todas as somas parciais o mesmo expoente de s) então devemos trocar os últimos p dígitos de s_k por zeros. Estes dígitos infuenciam os últimos dígitos de s, e, como em geral, estão errados, não podemos falar que o erro final será pequeno.

A perda de algarismos significativos devido a uma soma intermediária grande é chamada de **Propagação do Erro**. Veremos este fato através de exemplos.

Exemplo 2.15 - *Calcular $e^{-5.25}$, utilizando 5 dígitos significativos em todas as operações.*

Solução: O seguinte resultado matemático é bem conhecido: para todo número real x,

$$e^{-x} = \sum_{k=0}^{\infty} (-1)^k \frac{x^k}{k!} \ .$$

Se e^{-x} é calculado usando esta fórmula, a série deve ser truncada. Assim já estaremos introduzindo um erro de truncamento.

Vamos considerar os primeiros 20 termos da série acima para avaliar $e^{-5.25}$. Temos então:

$$\begin{aligned}
e^{-5.25} &= (0.10000 - 0.52500)10^1 + (0.13781 - 0.24117 + 0.31654 - 0.33236 \\
&+ 0.29082 - 0.21811 + 0.14314)10^2 + (-0.83497 + 0.43836 - 0.20922)10^1 \\
&+ (0.91532 - 0.36965 + 0.13862)10^0 + (-0.48516 + 0.15919)10^{-1} \\
&+ (-0.49164 + 0.14339)10^{-2} + (-0.39620 + 0.10401)10^{-3} \\
&+ (-0.26003)10^{-4} + (0.62050 - 0.14163)10^{-5} + (0.30982)10^{-6} \ .
\end{aligned}$$

Efetuando os cálculos obtemos: $e^{-5.25} = 0.65974 \times 10^{-2}$. Observe que, usando uma calculadora, o resultado de $e^{-5.25}$ é 0.52475×10^{-2}. Essa diferença entre os valores obtidos ocorreu porque na expressão acima temos parcelas da ordem de 10^2 que desprezam toda grandeza inferior a 10^{-3} (ver Tabela 2.1), enquanto que o resultado real de $e^{-5.25}$ é constituido quase que exclusivamente de grandezas dessa ordem. A pergunta que surge naturalmente é: podemos obter um resultado mais preciso? A resposta é sim. Basta lembrar que $e^{-5.25} = \frac{1}{e^{5.25}}$ e que:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} .$$

para todo número real x. Somando todas as parcelas da expressão de $e^{5.25}$, (desde que a expansão de e^x e e^{-x} diferem apenas em termos de sinal), obtemos: $e^{5.25} = 0.19057 \times 10^3$, e assim $e^{-5.25} = \frac{1}{e^{5.25}} = \frac{1}{0.19057 \times 10^3} = 0.52475 \times 10^{-2}$.

Na Tabela 2.1, apresentamos os cálculos de $e^{-5.25}$, $e^{-5.25}$ e $\frac{1}{e^{5.25}}$, considerando a expansão até o termo de ordem 10^k , $k = 1, 0, -1, \ldots, -6$.

Tabela 2.1

10^k	$e^{-5.25}$	$e^{5.25}$	$\frac{1}{e^{5.25}}$
10^1	$0.64130(10^0)$	$0.18907(10^3)$	$0.52890(10^{-2})$
10^0	$0.42990(10^{-1})$	$0.19049(10^3)$	$0.52496(10^{-2})$
10^{-1}	$0.10393(10^{-1})$	$0.19056(10^3)$	$0.52477(10^{-2})$
10^{-2}	$0.69105(10^{-2})$	$0.19056(10^3)$	$0.52477(10^{-2})$
10^{-3}	$0.66183(10^{-2})$	$0.19057(10^3)$	$0.52475(10^{-2})$
10^{-4}	$0.65929(10^{-2})$	$0.19057(10^3)$	$0.52475(10^{-2})$
10^{-5}	$0.65971(10^{-2})$	$0.19057(10^3)$	$0.52475(10^{-2})$
10^{-6}	$0.65974(10^{-2})$	$0.19057(10^3)$	$0.52475(10^{-2})$

Exemplo 2.16 - *Deseja-se determinar numericamente o valor exato da integral:*

$$y_n = \int_0^1 \frac{x^n}{x+a} \, dx ,$$

para um valor fixo de $a >> 1$ *e,* $n = 0, 1, \ldots, 10$.

Solução: Sabemos que os números y_n são positivos. Além disso, como para $0 < x < 1$, $x^{n+1} < x^n$, os números y_n formam uma sequência monotonicamente decrescente, e ainda:

$$\int_0^1 \frac{x^n}{1+a} \, dx < y_n < \int_0^1 \frac{x^n}{a} \, dx ,$$

e portanto podemos afirmar que:

$$\frac{1}{(n+1)(1+a)} < y_n < \frac{1}{(n+1)a} .$$

Assim para $a = 10$ e $n = 10$, temos que:

$$0.0082645 < y_{10} < 0.0090909 \ . \tag{2.2}$$

Agora para calcular numericamente o valor da integral dada, podemos primeiramente expressar o integrando usando o Teorema binomial, isto é:

$$x^n = [(x+a)-a]^n = \sum_{k=0}^{n} (-1)^k \begin{pmatrix} n \\ k \end{pmatrix} (x+a)^{n-k} a^k \ .$$

Substituindo x^n na expressão para y_n, obtemos:

$$\begin{aligned} y_n &= \int_0^1 \sum_{k=0}^{n} (-1)^k \begin{pmatrix} n \\ k \end{pmatrix} (x+a)^{n-k-1} a^k \ dx \\ &= \sum_{k=0}^{n} (-1)^k a^k \begin{pmatrix} n \\ k \end{pmatrix} \int_0^1 (x+a)^{n-k-1} \ dx \\ &= \sum_{k=0}^{n-1} (-1)^k a^k \begin{pmatrix} n \\ k \end{pmatrix} \int_0^1 (x+a)^{n-k-1} \ dx \\ &+ (-1)^n a^n \begin{pmatrix} n \\ n \end{pmatrix} \int_0^1 (x+a)^{-1} \ dx \end{aligned}$$

e assim:

$$\begin{aligned} y_n &= \sum_{k=0}^{n-1} (-1)^k a^k \begin{pmatrix} n \\ k \end{pmatrix} \left[\frac{1}{n-k} ((1+a)^n - k - a^{n-k}) \right] \\ &+ (-a)^n \ln \frac{1+a}{a}. \end{aligned} \tag{2.3}$$

Para $a = 10$ e $n = 10$, utilizando (2.3) para calcular y_n, obtemos $y_n = -2.000000$, que comparado com (2.2), está totalmente errado. A razão para isso é uma extrema propagação de erro. Para $n = 10$, o termo correspondente a $k = 5$ na soma (2.3) é igual a:

$$(-1)^5 a^5 \begin{pmatrix} 10 \\ 5 \end{pmatrix} \left[\frac{1}{5} ((1+a)^5 - a^5) \right] = -3.13 \times 10^{11} \ .$$

Assim para uma calculadora com 10 dígitos na mantissa, no mínimo dois dígitos antes do ponto decimal do resultado são não confiáveis, bem como os dígitos depois do ponto decimal.

2.5.3 Instabilidade Numérica

Se um resultado intermediário de um cálculo é contaminado por um erro de arredondamento, este erro pode influenciar todos os resultados subsequentes que dependem desse resultado intermediário. Os erros de arredondamento podem propagar-se mesmo que todos os cálculos subsequentes sejam feitos com precisão dupla. Na realidade, cada novo resultado intermediário introduz um novo erro de arredondamento. É de se esperar portanto, que todos esses erros influenciem o resultado final. Numa situação simples como o caso de uma soma, o erro final pode ser igual a soma de todos os erros intemediários.

Entretanto, os erros intermediários podem, algumas vezes, cancelar-se uns com os outros no mínimo parcialmente. Em outros casos (tal como em processos iterativos) os erros intermediários podem ter um efeito desprezível no resultado final. Algoritmos com essa propriedade são chamados **estáveis**.

Instabilidade Numérica ocorre se os erros intermediários tem uma influência muito grande no resultado final. Veremos esse fato através do seguinte exemplo.

Exemplo 2.17 - *Resolver a integral:*

$$I_n \;=\; e^{-1} \int_0^1 x^n e^x \; dx \; .$$

Solução: Vamos tentar encontrar uma fórmula de recorrência para I_n. Integrando por partes, segue que:

$$\begin{aligned} I_n &= e^{-1}\left\{ [x^n e^x]_0^1 - \int_0^1 \; n \; x^{n-1} e^x \; dx \right\} \\ &= 1 - n \; e^{-1} \int_0^1 \; x^{n-1} e^x \; dx \\ &= 1 - n \; I_{n-1} \; . \end{aligned}$$

Assim, obtemos uma fórmula de recorrência para I_n, isto é:

$$I_n \;=\; 1 - n \; I_{n-1} \; , \quad n = 1, 2, \ldots \; , \tag{2.4}$$

e desde que:

$$I_0 \;=\; e^{-1} \int_0^1 \; e^x \; dx \;=\; e^{-1}(e-1) \;=\; 0.6321,$$

é conhecido, podemos, teoricamente, calcular I_n, usando (2.4). Fazendo os cálculos, obtemos:

$$\begin{aligned} I_0 &= 0.6321 \; , \; I_1 \;=\; 0.3679 \; , \; I_2 \;=\; 0.2642 \; , \; I_3 \;=\; 0.2074 \; , \\ I_4 &= 0.1704 \; , \; I_5 \;=\; 0.1480 \; , \; I_6 \;=\; 0.1120 \; , \; I_7 \;=\; 0.216 \; . \end{aligned}$$

O resultado obtido para I_7 está claramente errado, desde que:

$$I_7 < e^{-1} \max_{0 \le x \le 1} (e^x) \; \int_0^1 x^n \; dx < \frac{1}{n+1} \; ,$$

isto é, $I_7 < \frac{1}{8} = 0.1250$. Além disso a sequência I_n é uma sequência decrescente. Para ver que a instabilidade existe, vamos supor que o valor de I_0 esteja afetado de um erro ϵ_0. Vamos supor ainda que todos as operações aritméticas subsequentes são calculadas exatamente. Denotando por I_n o valor exato da integral e por $\tilde{I}_n$ o valor calculado assumindo que só existe erro no valor inicial, obtemos que:

$$\tilde{I}_0 \; = I_0 + \epsilon_0 \; ,$$

e assim:

$$\tilde{I}_n \;=\; 1 - n \; \tilde{I}_{n-1} \; , \;\; n = 1, 2, \ldots \; . \tag{2.5}$$

Seja r_n o erro, isto é:

$$r_n \;=\; \tilde{I}_n - I_n \; .$$

Subtraindo (2.4) de (2.5), segue que:

$$r_n = -n\, r_{n-1} \ , \quad n = 1, 2, \ldots .$$

Aplicando essa fórmula repetidamente, obtemos:

$$r_n = -n r_{n-1} = (-n)^2 r_{n-2} = \ldots = (-n)^n r_0 \ ,$$

e portanto

$$r_n = (-n)^n \, \epsilon_0 \ ,$$

desde que $r_0 = \epsilon_0$. Assim, a cada passo do cálculo, o erro cresce do fator n. Surge então a pergunta: Como encontrar o valor exato de I_n? Para este caso em particular, observe que: *uma relação de recorrência ser instável na direção crescente de n não impede de ser estável na direção decrescente de n.* Assim, resolvendo (2.4), para I_{n-1}, obtemos:

$$I_{n-1} = \frac{(1 - I_n)}{n} \ . \tag{2.6}$$

Se usada nessa forma, a relação também precisa de um valor inicial. Entretanto, não é fácil encontrar esse valor pois todo I_n onde $n > 0$ é desconhecido. Mas sabemos que $I_n \to 0$ quando $n \to \infty$. Assim, tomando $I_{20} = 0$ e usando (2.6) para $n = 20, 19, 18, \ldots$, obtemos: $I_7 = 0.1123835$ onde agora todos os dígitos estão corretos. É interessante notar que começando com $I_7 = 0$, obtemos $I_0 = 0.6320$. Isto ocorre porque neste caso o erro está sendo reduzido substancialmente a cada passo, isto é, a cada passo o erro decresce do fator $\frac{1}{n}$.

2.5.4 Mal Condicionamento

A maioria dos processos numéricos seguem a seguinte linha geral:

- Dados são fornecidos,
- Os dados são processados de acordo com um plano pré-estabelecido (algoritmo),
- Resultados são produzidos.

Analisaremos aqui problemas onde os resultados dependem continuamente dos dados. Tais problemas são chamados de **problema bem posto**. Problemas que não dependem continuamente dos dados são chamados de **problema mal posto**.

Vamos então analisar como pertubações nos dados podem ou não influenciar os resultados.

Exemplo 2.18 - *Resolver o sistema:*

$$\begin{cases} x + y = 2 \\ x + 1.01y = 2.01 \end{cases}$$

Solução: A solução desse sistema pode ser facilmente obtida, por exemplo, por substituição. Fazendo isso, obtemos: x = y = 1. Se o número 2.01, da segunda equação é mudado para 2.02, obtemos que a solução do sistema é agora x=0 e y=2. Portanto uma pequena mudança nos dados produz uma grande mudança no resultado. Vamos então interpretar geometricamente o resultado. A solução do sistema é o ponto de interseção das duas retas: y = 2-x e y = (2.01 -x)/1.01. Essas retas estão desenhadas na Figura 2.1. É claro que o ponto de interseção é muito sensível a pequenas pertubações em cada uma dessas retas desde que elas são praticamente paralelas. De fato, se o coeficiente de y na segunda equação é 1.00, as duas retas são exatamente paralelas e o sistema não tem solução. Isto é típico de problemas mal condicionados. Eles são também chamados de problemas *críticos*, pois ou possuem infinitas soluções ou não possuem nenhuma.

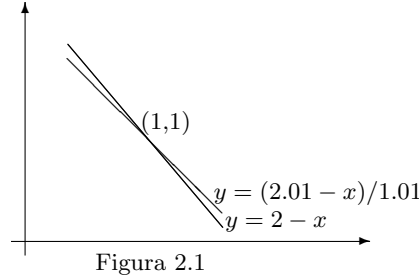



Figura 2.1

Exemplo 2.19 - *Determinar a solução do problema de valor inicial:*

$$\begin{cases} y'' & = & y \\ y(0) & = & a \\ y'(0) & = & b \end{cases}$$

onde a e b são dados.

Solução: A solução teórica desse problema de valor inicial é:

$$y(x) \; = \; C_1 e^x + C_2 e^{-x} \; , \tag{2.7}$$

onde C_1 e C_2 dependem de a e b. Assim, se tomarmos $a = 1$ e $b = -1$, então desde que $y'(x) = C_1 e^x - C_2 e^{-x}$, obtemos o sistema linear:

$$\begin{cases} y(0) \; = \; C_1 + C_2 & = & 1 \\ y'(0) \; = \; C_1 - C_2 & = & -1 \end{cases}$$

cuja solução é: $C_1 = 0$ e $C_2 = 1$. Substituindo esses valores em (2.7) obtemos que $y(x) = e^{-x}$. Logo, quando $x \to \infty$ a solução decresce rapidamente para zero. Mas se tomarmos $a = 1$ e $b = -1 + \delta$, onde $|\delta|$ pode ser arbitrariamente pequeno, então, como anteriormente, obtemos o sistema linear:

$$\begin{cases} C_1 + C_2 & = & 1 \\ C_1 - C_2 & = & -1 + \delta \end{cases}$$

cuja solução é: $C_1 = \frac{\delta}{2}$ e $C_2 = 1 - \frac{\delta}{2}$. Assim a solução do novo problema de valor inicial é:

$$y(x) = \frac{\delta}{2} e^x + (1 - \frac{\delta}{2}) e^{-x} = e^{-x} + \frac{\delta}{2}(e^x - e^{-x}) = e^{-x} + \delta \; senh \; x \; .$$

Portanto a solução difere da solução do problema anterior de δ *senh* x. Assim a característica matemática da solução foi mudada completamente, pois enquanto no primeiro resultado a solução $\to 0$, quando $x \to \infty$ ela agora $\to \infty$ quando $x \to \infty$. Tudo isso ocorreu apesar da dependência de $y(x)$ sobre os dados a e b ser claramente contínua.

Torna-se então necessário introduzir uma medida para o grau de continuidade de um problema. Tal medida é essencial em muitas definições de continuidade. Seja X o espaço dos dados; os elementos x de X podem ser números, pontos de um espaço euclidiano, vetores, matrizes, funções, etc.... Podemos então falar em continuidade se pudermos ser capazes de medir a distância entre os elementos de X. Suponhamos que o espaço X está dotado com uma função distância $d(x, y)$ que mede a distância entre os elementos x e y de X. Se por exemplo, X é o espaço dos números reais, a função distância é definida por: $d(x, y) = |x - y|$. Para $X = I\!R^n$, veja Definição 1.7.

Seja P o processo nos quais os dados x são transformados no resultado y, isto é: $y = P(x)$. Se o processo P é contínuo num ponto x, então a definição de continuidade (matemática) exige que para cada $\epsilon > 0$, $\exists\ \delta(\epsilon) > 0$ tais que:

$$|P(\tilde{x}) - P(x)| < \epsilon \ \text{ sempre que } \ |\tilde{x} - x| < \delta(\epsilon) \ .$$

Quanto maior a função $\delta(\epsilon)$ pode ser escolhida, mais contínuo é o processo P. No caso em que grandes mudanças nos dados produzem somente pequenas mudanças nos resultados, ou se $\delta(\epsilon)$ pode ser escolhido grande, a condição do problema é boa, e o problema é chamado **bem condicionado**. Por outro lado, se pequenas mudanças nos dados produzem grandes mudanças nos resultados, ou se $\delta(\epsilon)$ deve ser escolhido pequeno, a condição do problema é má, e o problema é chamado **mal condicionado**.

Exemplo 2.20 - *Analisar o problema de valor inicial do exemplo 2.19.*

Solução: Se queremos que a solução $y(x)$ num ponto x seja mudada por não mais que uma quantidade ϵ, então a condição inicial $y'(0) = -1$ deve ser mudada por não mais que: $\delta(\epsilon) = \frac{\epsilon}{senh\ x}$, o qual pode ser feito arbitrariamente pequeno escolhendo x grande. Por exemplo para $x = 10$, obtemos: $\delta(\epsilon) = 0.9 \times 10^{-4}\ \epsilon$. Assim temos um problema mal condicionado.

Podemos também verificar se um problema é ou não mal condicionado analisando o **número de condição** do problema. O problema será bem condicionado se o número de condição for pequeno e será mal condicionado se o número de condição for grande. Entretanto a definição de número de condição depende do problema.

Seja $y = P(x)$, com P diferenciável. Então a mudança em y causada pela mudança em x pode ser aproximada, (no sentido do cálculo diferencial) pelo diferencial de y, isto é: $dy = P'(x)\ dx$. Assim o comprimento de $|P'(x)|$ do operador linear $P(x)$ representa o número de condição do problema num ponto x. O número de condição relativa é definido por:

$$c_r \ = \ \frac{|P'(x)|}{|P(x)|}| \ .$$

Assim se $c_r \leq 1$ dizemos que o problema é relativamente bem condicionado.

Exemplo 2.21 - *Analisar o problema de calcular:*

$$f(x) \ = \ \left(\ln\ \frac{1}{x}\right)^{-\frac{1}{8}} \ ,$$

num ponto x qualquer.

Solução: Desde que f é diferenciável o número de condição é simplesmente $|f'(x)|$. Assim:

$$\begin{aligned} f'(x) \ &= \ -\frac{1}{8}\left(\ln\ \frac{1}{x}\right)^{-\frac{9}{8}} \frac{-1/x^2}{1/x} \\ &= \ \frac{1}{8\ x}\ \left(\ln\ \frac{1}{x}\right)^{-\frac{9}{8}} \ , \end{aligned}$$

e o número de condição relativa é dada por:

$$c_r = \left| \frac{f'(x)}{f(x)} \right| = \frac{1}{8\ x\ \ln\ \frac{1}{x}} \ .$$

Para $x = 0$, e $x = 1$ tanto o número de condição como o número de condição relativa são infinito, e assim nestes pontos o problema é extremamente mal condicionado. Para aproximadamente $0.1537 \leq x \leq 0.5360$, $c_r \leq 1$. Portanto neste intervalo o problema de calcular f é bem condicionado.

O problema de resolver um sistema linear, como vimos, é um outro exemplo de problema onde pequenas pertubações nos dados podem alterar de modo significativo o resultado. A análise do problema de mal condicionamento de sistemas lineares, encontra-se no Capítulo 4.

Teoricamente, o termo mal condicionado é usado somente para modelos matemáticos ou problemas e o termo instabilidade somente para algoritmos. Entretanto, na prática os dois termos são usados sem distinção.

O leitor interessado em maiores detalhes sobre os tópicos apresentados nesse capítulo, deve consultar os livros: [Forsythe, 19 ..] e [Henrice, 1982].

Exercícios

2.13 - *Considere a integral do exercício 2.13, com $a = 10$.*

a) *Calcule y_0 usando a integral.*

b) *Mostre que uma relação de recorrência para y_n é dada por:*

$$y_n = \frac{1}{n} - a\ y_{n-1} \ .$$

c) *Calcule $y_n,\ n = 1, 2 \ldots, 10$, usando a relação de recorrência.*

Os valores obtidos são confiáveis?

2.14 - *Considere agora a relação de recorrência do exercício anterior escrita na forma:*

$$y_{n-1} = \frac{1}{a} - \left(\frac{1}{n} - y_n \right) \ .$$

Considere ainda que $y_{20} = 0$. Usando este dado e a relação de recorrência obtenha os valores de $y_{10}, y_9, \ldots, y_1$. Os resultados agora são melhores? Como você explica isso?

2.6 Exercícios Complementares

2.15 - *Dados os números:* $(13.44)_5$, $(122.35)_6$, $(31.202)_4$. *Existe algum com representação exata no sistema $F(2, 10, 10, 10)$?*

2.16 - *Considere o sistema $F(2, 8, 4, 4)$ e os números $x_1 = 0.10110011 \times 2^2$ e $x_2 = 0.10110010 \times 2^2$. Qual dos dois números representa melhor $(2.8)_{10}$?*

2.17 - *Seja o sistema $F(2, 3, -1, 2)$. Exiba todos os números representáveis neste sistema e coloque-os sobre um eixo ordenado.*

2.18 - *Considere o sistema* $F(2,8,10,10)$*. Represente no sistema os números:* $x_1 = \sqrt{8}$, $x_2 = e^2$, $x_3 = 3.57$, *onde todos estão na base 10. Existe algum com representação exata nesse sistema?*

2.19 - *Mostre que se* x *é um número no sistema* $F(\beta,t,m,M)$ *então* $\bar{x} = x(1+\delta)$ *onde* $|\delta| \leq \frac{1}{2}\beta^{1-t}$.

2.20 - *Mostre que* $\frac{1}{2}\beta^{1-t}$ *é o melhor limitante para* $|\delta|$.

2.21 - *Efetue as operações indicadas, utilizando aritmética de ponto flutuante com 3 algarismos significativos.*

a) $(19.3 - 1.07) - 10.3$ *e* $19.3 - (1.07 + 10.3)$,

b) $27.2 \times 1.3 - 327. \times 0.00251$,

c) $\dfrac{10.1 - 3.1 \times 8.2}{14.1 + 7.09 \times 3.2^2}$,

d) $(367. + 0.6) + 0.5$ *e* $367. + (0.6 + 0.5)$,

e) $\sum_{i=1}^{100} 0.11$. *(Compare seu resultado com* 100×0.11*)* .

2.22 - *Deseja-se calcular:*

$$S = \sum_{k=1}^{10} \frac{2}{k^2} ,$$

no sistema $F(10,3,5,4)$*, usando arredondamento em todas as operações. Assim, efetue a soma:*

a) *da direita para a esquerda,*

b) *da esquerda para a direita,*

Os valores obtidos em **a)** *e* **b** *são iguais?*

2.23 - *Usando arredondamento para 4 dígitos significativos, efetue as operações indicadas, e escreva o resultado na forma normalizada.*

a) $0.5971 \times 10^3 + 0.4268 \times 10^0$,

b) $0.5971 \times 10^{-1} - 0.5956 \times 10^{-2}$,

c) $\dfrac{0.5971 \times 10^3}{0.4268 \times 10^{-1}}$,

d) $(0.5971 \times 10^3) \times (0.4268 \times 10^0)$,

2.24 - *Usando arredondamento, a cada operação efetuada, calcule:*

$$\sum_{i=1}^{n} i^2 - \sum_{i=2}^{n} i^2 = 1 ,$$

somando os termos em:

a) *ordem crescente,*

b) *ordem decrescente.*

Considere $n = 100$. *Os valores obtidos são iguais?*

2.25 - *Deseja-se calcular* $e^{-0.15}$.

a) *Obtenha através de uma calculadora o valor exato de* $e^{-0.15}$.

b) *Considere o sistema* $F(10, 5, 10, 10)$ *e a série truncada em 25 termos. Calcule:*

b.1) $e^{-0.15}$

b.2) $\frac{1}{e^{0.15}}$

e compare os resultados.

2.26 - *Se* a, b *e* c *são reais e* $a \neq 0$, *então a equação*

$$a\,x^2 \;+\; b\,x \;+\; c \;=\; 0\;,$$

é satisfeita para exatamente dois valores de x:

$$(I)\begin{cases} x_1 \;=\; \dfrac{-b \;+\; \sqrt{b^2\;-4\;a\;c}}{2\;a}\;, \\[2ex] x_2 \;=\; \dfrac{-b \;-\; \sqrt{b^2\;-\;4\;a\;c}}{2\;a}\;. \end{cases}$$

Entretanto, desde que, $a\;x^2 + b\;x + c = a\;(x - x_1)\;(x - x_2)$ *obtemos que:* $a\;x_1\;x_2 \;=\; c$ *e assim podemos reescrever (I) na seguinte forma:*

$$(II)\begin{cases} x_1 \;=\; -\dfrac{b \;+\; \text{sinal de }(b)\;\sqrt{b^2\;-\;4\;a\;c}}{2\;a}\;, \\[2ex] x_2 \;=\; \dfrac{c}{a\;x_1}\;, \end{cases}$$

Temos ainda que x_1 *e* x_2 *podem ser escritos como:*

$$(III)\begin{cases} x_1 \;=\; \dfrac{-2\;c}{b \;+\; \sqrt{b^2\;-\;4\;a\;c}}\;, \\[2ex] x_2 \;=\; \dfrac{c}{a\;x_1}\;. \end{cases}$$

Utilizando (I), (II) e (III) calcule as raízes das equações para os valores de a, b *e* c *dados a seguir:*

i) $a \;=\; 1\;;\quad b \;=\; -10^5\;;\quad c \;=\; 1\;\;,$

ii) $a \;=\; 1\;;\quad b \;=\; -4\;;\quad c \;=\; 3.9999999\;\;,$

iii) $a \;=\; 6\;;\quad b \;=\; 5\;;\quad c \;=\; -4\;\;.$

2.27 - *A função de Bessel satisfaz a seguinte relação de recorrência:*

$$J_{n+1}(x) - \frac{2\;n}{x} J_n(x) + J_{n-1}(x) = 0\;.$$

Se $x = 1$, $J_0(1) = 0.7652$ *e* $J_1(1) = 0.4401$, *calcule* $J_n(1)$ *para* $n = 2, 3, \ldots, 10$. *Refaça os cálculos começando com valores mais precisos:* $J_0(1) = 0.76519769$ *e* $J_1(1) = 0.44005059$. *Como você explica seus resultados com o fato que* $J_n(1) \to 0$ *quando* n *cresce?*

2.28 - *Faça* $J_{10}(1) = 0$ *e* $J_9(1) = \mu$. *Use a fórmula do exercício anterior, na forma:*

$$J_{n-1}(1) \; = \; 2 \; nJ_n(1) - J_{n+1}(1) \; .$$

e calcule $J_8(1), J_7(1), \ldots$. *Encontre* μ *através da identidade:*

$$J_0(x) + 2 \; J_2(x) + 2 \; J_4(x) + 2 \; J_6(x) + \ldots \; = \; 1 \; ,$$

e calcule $J_9(1), J_8(1), \ldots, J_0(1)$. *Como esses resultados se comparam com os valores exatos?*

Capítulo 3

Equações não Lineares

3.1 Introdução

Um dos problemas que ocorrem mais frequentemente em trabalhos científicos é calcular as raízes de equações da forma:

$$f(x) = 0 ,$$

onde $f(x)$ pode ser um polinômio em x ou uma função transcendente. Em raros casos é possível obter as raízes exatas de $f(x) = 0$, como ocorre por exemplo, supondo-se $f(x)$ um polinômio fatorável. Através de técnicas numéricas, é possível obter uma solução aproximada, em alguns casos, tão próxima da solução exata, quanto se deseje. A maioria dos procedimentos numéricos fornecem uma sequência de aproximações, cada uma das quais mais precisa que a anterior, de tal modo que a repetição do procedimento fornece uma aproximação a qual difere do valor verdadeiro por alguma tolêrancia pré-fixada. Estes procedimentos são portanto muito semelhantes ao conceito de limite da análise matemática. Vamos considerar vários **métodos iterativos** para a determinação de aproximações para raízes isoladas de $f(x) = 0$. Será dada uma atenção especial às equações polinomiais em virtude da importância que as mesmas gozam em aplicações práticas.

Inicialmente recordemos um importante resultado da Álgebra.

Teorema 3.1 - *Se uma função contínua $f(x)$ assume valores de sinais opostos nos pontos extremos do intervalo $[a, b]$, isto é, se $f(a) \times f(b) < 0$, então existe pelo menos um ponto $\bar{x} \in [a, b]$, tal que $f(\bar{x}) = 0$.*

Prova: A prova deste teorema pode ser encontrada em [.. , 19..]

Definição 3.1 - *Se $f : [a, b] \to I\!R$ é uma função dada, um ponto $\bar{x} \in [a, b]$ é um zero (ou raiz) de f se $f(\bar{x}) = 0$.*

Ilustraremos graficamente esses conceitos nos exemplos a seguir.

Exemplo 3.1 - *Seja $f : (0, \infty) \to I\!R$. Determinar as raízes de $f(x) = \ln\ x$.*

Solução: O gráfico de $\ln\ x$ é dado na Figura 3.1.

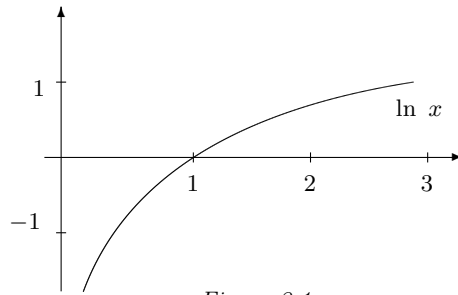



Figura 3.1

Nesse caso vemos que $f(0.5) \times f(1.5) < 0$. Portanto existe uma raiz de $f(x)$ no intervalo $(0.5, 1.5)$. Além disso a curva intercepta o eixo dos x num único ponto, pois trata-se de uma função crescente. Então $\bar{x} = 1$ é a única raiz de $f(x) = 0$.

Exemplo 3.2 - *Seja* $f : (0, \infty) \rightarrow I\!R$*. Determinar as raízes de* $f(x) = e^x$*.*

Solução: O gráfico de e^x é dado na Figura 3.2.

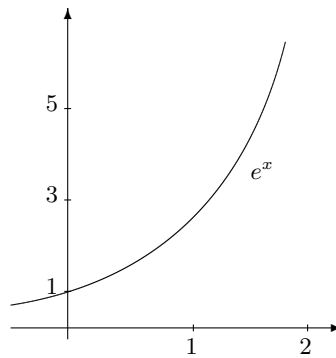



Figura 3.2

Nesse caso vemos a curva não intercepta o eixo dos x, logo não existe $\bar{x}$ tal que $f(\bar{x}) = 0$.

Exemplo 3.3 - *Seja* $f : [0, 2\pi] \rightarrow I\!R$*. Determinar as raízes de* $f(x) = cos\ x$*.*

Solução: O gráfico de *cos x* é dado na Figura 3.3.

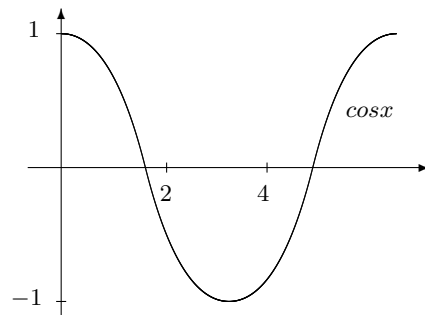



Figura 3.3

Nesse caso vemos que: $f(1) \times f(2) < 0$ e $f(4) \times f(5) < 0$, ou seja a curva intercepta o eixo dos x em dois pontos. Assim temos uma raiz $\bar{x}$ no intervalo $(1, 2)$ e outra no intervalo $(4, 5)$. Sabemos da trigonometria que: $\bar{x} = \frac{\pi}{2} \simeq 1.5708$ e $\bar{x} = \frac{3\,\pi}{2} \simeq 4.7124$ são raízes de $f(x) = 0$.

Definição 3.2 - *Um ponto* $\bar{x} \in [a, b]$ *é uma raiz de multiplicidade* m *da equação* $f(x) = 0$ *se* $f(x) = (x - \bar{x})^m \; g(x)$*; com* $g(\bar{x}) \neq 0$ *em* $[a, b]$.

Exemplo 3.4 - *Seja* $f : I\!R \to I\!R$*. Determinar as raízes de* $f(x) = x^2 + 2\,x + 1 = (x + 1)^2 = 0$.

Solução: O gráfico de $f(x)$ é dado na Figura 3.4.

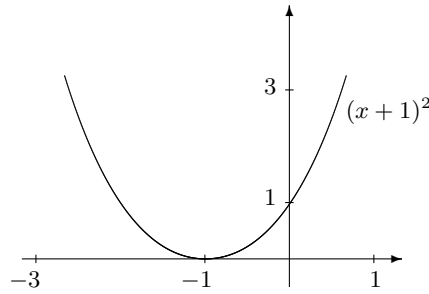



Figura 3.4

Nesse caso vemos que a curva apenas toca o eixo dos x. Assim, $\bar{x} = 1$ é raiz de multiplicidade 2 de $f(x) = 0$.

Como vimos nos exemplos anteriores, podemos obter o número exato de raízes e sua localização exata ou aproximada traçando o gráfico da função e encontrando o ponto onde a curva intercepta o eixo dos x. Entretanto algumas vezes é mais conveniente rearranjar a equação dada como $y_1(x) = y_2(x)$, para duas funções y_1 e y_2, cujos gráficos são mais fáceis de serem traçados do que o da f. As raízes da equação original são dadas então pelos pontos onde o gráfico de y_1 intercepta o de y_2. Ilustraremos este fato no próximo exemplo.

Exemplo 3.5 - *Seja* $f : I\!R \to I\!R$*. Determinar as raízes de* $f(x) = (x + 1)^2 \; e^{(x^2 - 2)} - 1 = 0$.

Solução: Podemos rearranjar a equação dada, por exemplo, como:

$$(x + 1)^2 = e^{(2 - x^2)}.$$

Fazendo $y_1 = (x + 1)^2$, $y_2 = e^{(2 - x^2)}$ e colocando as duas curvas no mesmo gráfico, obtemos a Figura 3.5.

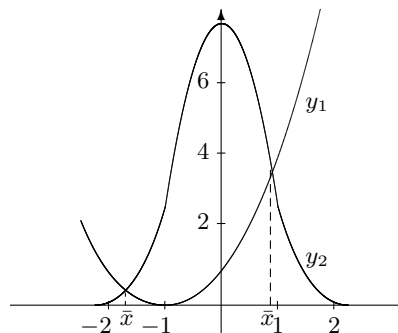



Figura 3.5

É claro observando-se a Figura 3.5 que as duas curvas se interceptam apenas duas vezes. Portanto a equação dada tem precisamente duas raízes. Uma raiz $\bar{x}$ no intervalo $(-2,-1)$ e outra no intervalo $(0,1)$.

Este último exemplo ilustra bem a razão da utilização de métodos númericos para determinar a solução de equações não lineares. Ao contrário dos exemplos anteriores, onde foi razoavelmente fácil determinar as raízes da função dada, aqui fica difícil dizer com exatidão qual é o valor de $\bar{x}$ tal que $f(\bar{x}) = 0$.

Para descrevermos um método numérico extremamente simples, e de fácil compreensão, suponha que $f(x)$ seja uma função contínua em $[a,b]$. Pelo Teorema 3.1, temos que se $f(x)$ em $x = a$ e $x = b$ tem sinais opostos, então $f(x)$ tem no mínimo um zero em $[a,b]$. Esse resultado fornece um caminho simples, mas efetivo, para encontrar a localização aproximada dos zeros da f. Considere novamente a equação do exemplo 3.5, isto é, $f(x) = (x+1)^2 \; e^{(x^2-2)} - 1$. Valores de $f(x)$ para $x = -3, -2, \ldots, 3$ estão contidos na tabela a seguir:

x	-3	-2	-1	0	1	2	3
$f(x)$	4385.5	6.4	−1.0	−0.9	0.5	65.5	17545.1

A função portanto possui zeros no intervalo $[-2,-1]$ e $[0,1]$. (Note que o mesmo resultado foi obtido graficamente). Estamos agora em condições de descrever um método numérico, conhecido como **Método da Bissecção**, o qual reduz o comprimento do intervalo que contém a raiz, de maneira sistemática.

Considere o intervalo $[a,b]$ para o qual $f(a) \times f(b) < 0$. No método da bissecção calculamos o valor da função $f(x)$ no ponto médio: $x_1 = \dfrac{a+b}{2}$. Portanto existem três possiblidades. Primeiramente, ficaríamos felizes, (embora seja quase impossível), se o valor da função calculado no ponto x_1 fosse nulo, isto é: $f(x_1) = 0$. Nesse caso x_1 é o zero da f e não precisamos fazer mais nada. Em segundo lugar, se $f(a) \times f(x_1) < 0$, então f tem um zero entre a e x_1. O processo pode ser repetido sobre o novo intervalo $[a, x_1]$. Finalmente, se $f(a) \times f(x_1) > 0$, segue que $f(b) \times f(x_1) < 0$, desde que é conhecido que $f(a)$ e $f(b)$ têm sinais opostos. Portanto f tem um zero entre x_1 e b, e o processo pode ser repetido com $[x_1, b]$. A repetição do método é chamado *iteração* e as aproximações sucessivas são os *termos iterados*. Assim, o método da bissecção pode ser descrito como:

Para $k = 1, 2, \ldots$, faça:

$$x_k = \frac{a+b}{2} \; .$$

$$\text{Se } f(a) \times f(x_k) \begin{cases} < 0 \quad \text{então} \quad b = x_k \; , \\ \\ > 0 \quad \text{então} \quad a = x_k \; \; . \end{cases}$$

Uma interpretação geométrica do método da bissecção é dada na Figura 3.6.

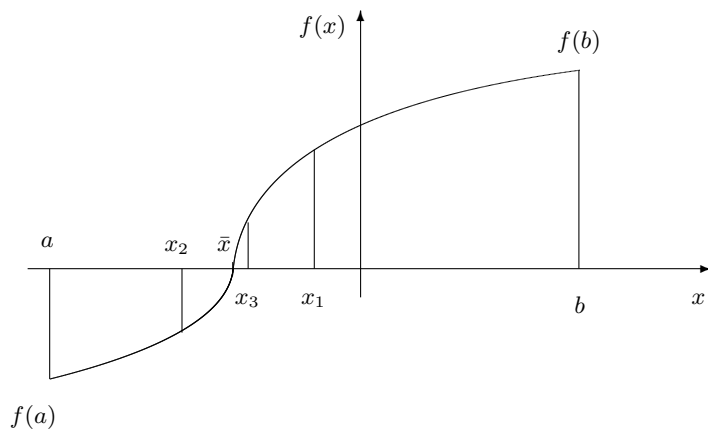



Figura 3.6

Para ilustrar o método da bissecção, considere que desejamos calcular a raiz positiva da equação do Exemplo 3.5, iniciando com o intervalo $[0, 1]$. Para essa equação temos que $f(0) < 0$ e $f(1) > 0$. O ponto médio é $x_1 = 0.5$, com $f(x_1) = -0.6090086$. Desde que $f(0) \times f(0.5) > 0$, deduzimos que a raiz da equação está em $[0.5, 1]$. Os primeiros passos do método da bissecção, para esta equação, estão mostrados na tabela:

k	a	b	x_k	$f(x_k)$
1	0	1	0.5	-0.609009
2	0.5	1	0.75	-0.272592
4	0.75	1	0.875	0.023105
4	0.75	0.875	0.8125	-0.139662
5	0.8125	0.875	0.84375	-0.062448
6	0.84375	0.875	0.859375	-0.020775
⋮				

Continuando o processo obteremos: $x_{16} = 0.866868$ e $x_{17} = 0.866876$. Isso significa que o intervalo incial $[0, 1]$ foi reduzido ao intervalo$[0.866868, 0.866876]$, e portanto a raiz positiva da equação dada é aproximadamente: $\bar{x} = 0.86687$. Note que até agora não falamos como devemos proceder para obter o resultado com uma quantidade de casas decimais corretas. Isso será discutido mais adiante.

Exercícios

3.1 - *Dadas as funções:*

a) $x^3 + 3\ x - 1 = 0,$

b) $x^2 - sen\ x = 0,$

pesquisar a existência de raízes reais e isolá-las em intervalos.

3.2 - *Justifique que a função:*

$$f(x) \;=\; cos\ \frac{\pi(x+1)}{8} + 0.148\ x - 0.9062\ ,$$

possui uma raiz no intervalo $(-1, 0)$ *e outra no intervalo* $(0, 1)$.

Existem vários métodos numéricos para determinação (aproximada) das raízes da equação $f(x) = 0$, mais eficientes que o método da bissecção. Descreveremos a seguir alguns desses métodos, discutindo suas vantagens e desvantagens. Antes, porém, daremos um procedimento que deve ser seguido na aplicação de qualquer método numérico para determinar um zero de $f(x) = 0$, com uma precisão pré-fixada.

Processo de Parada

1) Para aplicar qualquer método numérico deveremos ter sempre uma idéia sobre a localização da raiz a ser determinada. Essa localização é obtida , em geral, através de gráfico. (Podemos também localizar o intervalo que contém a raiz fazendo uso do Teorema 3.1). A partir da localização da raiz, escolhemos então x_0 como uma aproximação inicial para a raiz $\bar{x}$ de $f(x) = 0$. Com essa aproximação inicial e um método numérico *refinamos* a solução até obtê-la com uma determinada precisão (número de casas decimais corretas) .

2) Para obtermos uma raiz com uma determinada precisão ϵ devemos, durante o processo iterativo, efetuar o seguinte teste: Se

$$\frac{|x_{k+1} - x_k|}{|x_{k+1}|} < \epsilon \text{ , } (erro\ relativo) \text{ ,}$$

onde ϵ é uma precisão pré-fixada; x_k e x_{k+1} são duas aproximações consecutivas para $\bar{x}$, então x_{k+1} é a raiz procurada, isto é, tomamos $\bar{x} = x_{k+1}$.

Observações:

a) Em relação à precisão pré-fixada, normalmente tomamos $\epsilon = 10^{-m}$ onde m é o número de casas decimais que queremos corretas no resultado.

b) Apesar de alguns autores considerarem como teste de parada o fato de $|f(x_{k+1})| < \epsilon$, é preciso ter muito cuidado pois a menos que se tenha uma idéia muito clara do comportamento da função o fato desse teste ser satisfeito não implica necessariamente que x_{k+1} esteja próximo da raiz procurada, como pode ser observado no seguinte exemplo: considere $f(x) = x^{-3}\ lnx = 0$, onde a única raiz é $\bar{x} = 1$. Calculando $f(x)$ para $x = 2, 4, 8, 16, 32, \ldots$ obtemos, respectivamente: 0.0866, 0.0217, 0.00406,0.0006769,0.0001058,... isto é, quanto mais longe estamos de $\bar{x}$, menor é o valor de $f(x)$.

c) Alguns autores consideram como teste de parada o fato de $|x_{k+1} - x_k| < \epsilon$, chamado de **erro absoluto**. Entretanto, se esses números forem muito grandes e ϵ for muito pequeno, pode não ser possível calcular a raiz com uma precisão tão exigente. Como exemplo, resolva a equação $f(x) = (x - 1)(x - 2000) = 0$ com $\epsilon = 10^{-4}$ usando os critérios de erro relativo e erro absoluto. Você irá verificar que o número de iterações é muito maior para o critério do erro absoluto. Isso ocorre porque a raiz que estamos procurando tem módulo grande e portanto é muito mais difícil tornar o erro absoluto menor do que ϵ.

Quando fazemos um programa computacional, devemos considerar o erro relativo escrito na seguinte forma:

$$|x_{k+1} - x_k| < \epsilon * \max\{1, |x_{k+1}|\} \text{ ,}$$

pois se $|x_{k+1}|$ estiver próximo de zero o processo não estaciona. Além do teste do erro relativo, devemos colocar um número máximo de iterações pois se o programa não estiver bem, ou se o método não se aplicar ao problema que se está resolvendo, o programa entrará em *looping*.

3.2 Iteração Linear

A fim de introduzir o método de iteração linear para o cálculo de uma raiz da equação:

$$f(x) \;=\; 0 \;, \tag{3.1}$$

onde $f(x)$ é uma função contínua num intervalo que contenha a raiz procurada, expressamos, inicialmente, a equação (3.1) na forma:

$$x \;=\; \psi(x) \;, \tag{3.2}$$

de maneira que qualquer solução de (3.2) seja, também, solução de (3.1). Para qualquer função ψ, qualquer solução de (3.2) é chamada de **ponto fixo** de $\psi(x)$. Assim, o problema de determinar um zero de $f(x)$ foi transformado no problema de determinar o ponto fixo de $\psi(x)$, e essa transformação não deve alterar a posição da raiz procurada.

Em geral, há muitos modos de expressar $f(x)$ na forma (3.2). Basta considerarmos:

$$\psi(x) \;=\; x + A(x)\; f(x) \;,$$

para qualquer $A(x)$ tal que $A(\bar{x}) \;\neq\; 0$.

Nem todas, porém, serão igualmente satisfatórias para as nossas finalidades.

Algumas formas possíveis da equação:

$$f(x) \;=\; x^2 - x - 2 \;=\; 0; \tag{3.3}$$

cujas raízes são -1 e 2, por exemplo, são:

a) $x \;=\; x^2 - 2$ c) $x \;=\; 1 + \dfrac{2}{x}$

b) $x \;=\; \sqrt{2+x}$ d) $x \;=\; x \;-\; \dfrac{x^2 - 2x - 8}{m}, \; m \;\neq\; 0.$

É claro, que não necessitamos de um método númerico para calcular as raízes de uma equação do segundo grau, contudo esse exemplo ilustrará de maneira objetiva os nossos propósitos.

Como já dissemos anteriormente, geometricamente, (3.1) tem como solução a intersecção do gráfico da f com o eixo x, enquanto que uma raiz de (3.2) é um número $\bar{x}$, para o qual a reta $y_1 = x$ intercepta a curva $y_2 \;=\; \psi(x)$. Pode ocorrer, naturalmente, que estas curvas não se interceptem, caso em que não haverá raiz real. Admitiremos, contudo, que essas curvas se interceptem no mínimo, uma vez; que estamos interessados em determinar uma dessas raízes, digamos $\bar{x}$, e que $\psi(x)$ e $\psi'(x)$ sejam contínuas num intervalo que contenha essa raiz.

Seja x_0 uma aproximação inicial para a raiz $\bar{x}$ de (3.2). Obtemos as aproximações sucessivas x_k para a solução desejada $\bar{x}$, usando o processo iterativo definido por:

$$x_{k+1} \;=\; \psi\,(x_k)\,, \; k \;=\; 0, 1, \ldots . \tag{3.4}$$

Esse processo é chamado **Método Iterativo Linear**.

Para que esse processo seja vantajoso, devemos obter aproximações sucessivas x_k, convergentes para a solução desejada $\bar{x}$. Contudo, é fácil obter exemplos para os quais a sequência x_k diverge.

Exemplo 3.6 - *Considerando em (3.3),* $x \;=\; x^2 - 2$ *e tomando* $x_0 = 2.5$*, determinar a raiz* $\bar{x} = 2$*.*

Solução: Usando (3.4), obtemos:

$$\begin{aligned}
x_1 &= \psi\,(x_0) \;=\; x_0^2 \;-\; 2 \;=\; (2.5)^2 - 2 \;=\; 4.25 \\
x_2 &= \psi\,(x_1) \;=\; x_1^2 \;-\; 2 = \; (4.25)^2 - 2 \;=\; 16.0625 \\
x_3 &= \psi\,(x_2) \;=\; x_2^2 \;-\; 2 \;=\; (16.0625)^2 - 2 \;=\; 256.00391 \\
&\vdots
\end{aligned}$$

e é óbvio que se trata de uma sequência divergente. Assim, a escolha de $\psi(x) = x^2 - 2$ não produz um processo iterativo que seja convergente.

As **condições suficientes** que a função $\psi(x)$ deve satisfazer para assegurar a convergência da iteração linear estão contidas no Teorema 3.4. Vejamos antes dois teoremas que serão utilizados na prova desse Teorema.

Teorema 3.2 - **Teorema do Valor Médio** - *Se f é contínua em [a,b] e diferenciável em (a,b) então existe pelo menos um ponto ξ entre a e b tal que:*

$$f'(\xi) \;=\; \frac{f(b) - f(a)}{b - a} \;,\quad \text{isto é} \;,\quad f(b) - f(a) = f'(\xi)(b - a) \;.$$

Prova: A prova deste teorema pode ser encontrada em [Swokowski,1983].

Teorema 3.3 - **Teorema da Permanência do Sinal** - *Seja f uma função real de variável real definida e contínua numa vizinhança de x_0. Se $f(x_0) \neq 0$ então $f(x) \neq 0$ para todo x pertencente a um vizinhança suficientemente pequena de x_0.*

Prova: A prova deste teorema pode ser encontrada em [Swokowski,1983].

Teorema 3.4 - *Seja $\psi(x)$ uma função contínua, com derivadas primeira e segunda contínuas num intervalo fechado I da forma $I = (\bar{x} - h, \bar{x} + h)$, cujo centro $\bar{x}$ é solução de $x = \psi(x)$. Seja $x_0 \in I$ e M um limitante da forma, $|\psi'(x)| \leq M < 1$ em I. Então:*

a) *a iteração $x_{k+1} = \psi(x_k)$, $k = 0, 1, \ldots$, pode ser executada indefinidamente, pois $x_k \in I$, $\forall\ k$.*

b) $|x_k - \bar{x}| \to 0$.

c) *Se $\psi'(\bar{x}) \neq 0$ ou $\psi'(\bar{x}) = 0$ e $\psi''(\bar{x}) \neq 0$ e se $|x_0 - \bar{x}|$ for suficientemente pequeno então a sequência $x_1, x_2, \ldots$ será monotônica ou oscilante.*

Prova:

a) Usaremos indução para provar que $x_k \in I$, $\forall\ k$.

i) Por hipótese $x_0 \in I$.

ii) Supomos que $x_0, x_1, \ldots, x_k \in I$.

iii) Provemos que $x_{k+1} \in I$. Temos:

$$x_{k+1} - \bar{x} \;=\; \psi\left(x_k\right) \;-\; \psi\left(\bar{x}\right).$$

Usando o teorema do Valor Médio, obtemos:

$$x_{k+1} - \bar{x} \;=\; \psi'\left(\xi_k\right)\left(x_k - \bar{x}\right)$$

onde ξ_k está entre x_k e $\bar{x}$. Tomando módulo, segue que:

$$|x_{k+1} - \bar{x}| \;=\; |\psi'\left(\xi_k\right)||x_k - \bar{x}| \;\leq\; M|x_k - \bar{x}|$$

desde que pela hipótese de indução $x_k \in I \;\Rightarrow\; \xi_k \in I$ e sobre $I, |\psi'(x)| \leq M < 1$. Assim:

$$|x_{k+1} - \bar{x}| \;\leq\; M|x_k - \bar{x}|.$$

Como $M < 1$, temos que $x_{k+1} \in I$.

b) Pelo item a) temos que:

$$|x_k - \bar{x}| \leq M|x_{k-1} - \bar{x}| \leq M^2 |x_{k-2} - \bar{x}| \leq \ldots \leq M^k|x_0 - \bar{x}| .$$

Como $M < 1$, passando ao limite, obtemos:

$$\lim_{k\to\infty} M^k \to 0 \text{ e portanto } |x_k - \bar{x}| \to 0 .$$

c) Aqui dividiremos a prova em duas partes. Assim:

c.1) Seja $\psi'(\bar{x}) \neq 0$.

Pelo Teorema da Permanência do Sinal temos que numa vizinhança de $\bar{x}$ suficientemente pequena, $\psi'(x)$ terá o mesmo sinal. Assim de:

$$x_{k+1} - \bar{x} = \psi'(\xi_k)(x_k - \bar{x}) ,$$

temos que:

$$(I) \text{ Se } \psi'(\bar{x}) > 0 \text{ e } \begin{cases} x_k \leq \bar{x} \Rightarrow x_{k+1} \leq \bar{x} \\ x_k \geq \bar{x} \Rightarrow x_{k+1} \geq \bar{x} \end{cases}$$

$$(II) \text{ Se } \psi'(\bar{x}) < 0 \text{ e } \begin{cases} x_k \leq \bar{x} \Rightarrow x_{k+1} \geq \bar{x} \\ x_k \geq \bar{x} \Rightarrow x_{k+1} \leq \bar{x} \end{cases}$$

Como $|x_k - \bar{x}| \to 0$, a convergência será monotônica em (I) e em (II) será oscilante em torno de $\bar{x}$.

c.2) Seja $\psi'(\bar{x}) = 0$ e $\psi''(\bar{x}) \neq 0$.

Usando o teorema do Valor Médio, temos que:

$$\psi'(\xi_k) = \psi'(\xi_k) - \psi'(\bar{x}) = \psi''(\theta_k)(\xi_k - \bar{x}) ,$$

onde θ_k está entre ξ_k e $\bar{x}$. Assim:

$$x_{k+1} - \bar{x} = \psi''(\theta_k)(\xi_k - \bar{x})(x_k - \bar{x}) .$$

Pelo teorema da Permanência do Sinal, $\psi''(x)$ terá o mesmo sinal numa vizinhança suficientemente pequena de $\bar{x}$. Como $(\xi_k - \bar{x})(x_k - \bar{x}) \geq 0$, pois ξ_k e x_k encontram-se do mesmo lado de $\bar{x}$, segue que, se:

$\psi''(\bar{x}) > 0 \quad \Rightarrow \quad x_{k+1} \geq \bar{x}, \ \forall\, k$,

$\psi''(\bar{x}) < 0 \quad \Rightarrow \quad x_{k+1} \leq \bar{x}, \ \forall\, k$.

Neste caso a sequência $x_1, x_2, \ldots$ será monotônica independente do sinal de $x_0 - \bar{x}$. Isso completa a prova do Teorema 3.4.

Consideremos novamente a equação (3.3). Se nosso objetivo é encontrar a raiz $\bar{x} = 2$, usando o problema de ponto fixo equivalente (3.3.a), teremos:

$$x = x^2 - 2 = \psi(x).$$

Para que o processo $x_{k+1} = x_k^2 - 2$ seja convergente devemos ter $|\psi'(x)| < 1$ na vizinhança de $\bar{x} = 2$. Temos que $\psi'(x) = 2x$, e desde que $|\psi'(x)| > 1$ para $x > \frac{1}{2}$, o Teorema 3.4 não pode ser usado para garantir convergência. Entretanto, a iteração $x_{k+1} = x_k^2 - 2$ divergirá para qualquer escolha de $x_0 > \frac{1}{2}$, como vimos anteriormente.

Por outro lado se usarmos o problema de ponto fixo (3.3.b), teremos $\psi(x) = \sqrt{2+x}$, e assim $\psi'(x) = \frac{1}{2\sqrt{2+x}}$. Portanto $|\psi'(x)| < 1$, se e somente se $x > -1.75$. Assim, pelo Teorema 3.4, podemos crer que a iteração:

$$x_{k+1} \;=\; \sqrt{2+x_k}\ ,$$

será convergente para qualquer escolha de $x_0 > -1.75$, como pode ser observado no próximo exemplo.

Exemplo 3.7 - *Considerando em (3.3), $x = \sqrt{2+x}$ e tomando $x_0 = 2.5$, determinar a raiz $\bar{x} = 2$.*

Solução: Tomando $x_0 = 2.5$, obteremos a sequência de aproximações:

$$\begin{array}{lllllllll}
x_1 & = & \psi(x_0) & = & \sqrt{2+2.5} & = & \sqrt{4.5} & = & 2.1213203 \\
x_2 & = & \psi(x_1) & = & \sqrt{2+2.1213203} & = & \sqrt{4.1213203} & = & 2.0301035 \\
x_3 & = & \psi(x_2) & = & \sqrt{2+2.0301035} & = & \sqrt{4.0301035} & = & 2.0075118 \\
x_4 & = & \psi(x_3) & = & \sqrt{2+2.0075118} & = & \sqrt{4.0075118} & = & 2.0018771 \\
x_5 & = & \psi(x_4) & = & \sqrt{2+2.0018771} & = & \sqrt{4.0018771} & = & 2.0004692 \\
x_6 & = & \psi(x_5) & = & \sqrt{2+2.0004692} & = & \sqrt{4.0004692} & = & 2.0001173 \\
x_7 & = & \psi(x_6) & = & \sqrt{2+2.0001173} & = & \sqrt{4.0001173} & = & 2.0000293 \\
\vdots & & & & & & & &
\end{array}$$

a qual é, obviamente, convergente para a raiz $\bar{x} = 2$. Este exemplo ilustra, também, a importância da disposição apropriada de (3.1) na forma (3.2).

Uma ilustração geométrica da não convergência e da convergência do método iterativo $x_{k+1} = \psi(x_k)$ em ambos os casos: $x_{k+1} = x_k^2 - 2$ e $x_{k+1} = \sqrt{2+x}$ é dada pelas Figuras 3.7 e 3.8, respectivamente. Observe que em cada uma das figuras, escolhido o ponto x_0 caminhamos verticalmente até encontrar a curva $\psi(x)$; em seguida caminhamos horizontalmente até encontrar a reta $y = x$, e finalmente caminhamos verticalmente até encontra o eixo dos x onde estará localizado o ponto x_1. O processo é repetido partindo-se de x_1, e assim sucessivamente.Temos então:

a) para $x = x^2 - 2$:

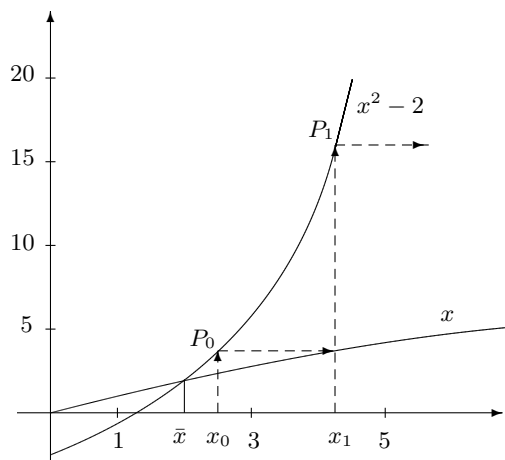



Figura 3.7

b) para $x = \sqrt{2+x}$:

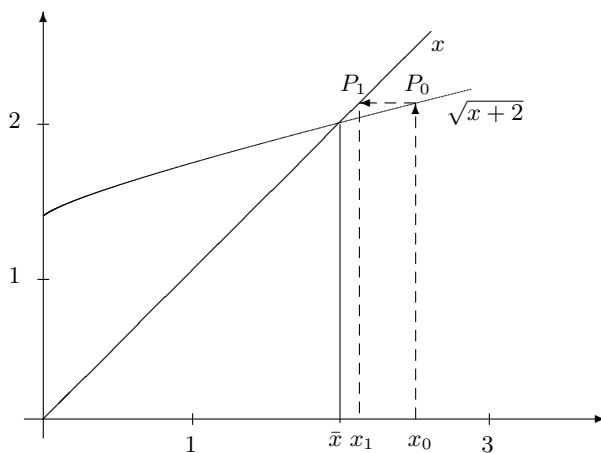



Figura 3.8

Representamos na Figura 3.7 os pontos: $P_0 : (x_0, \psi(x_0))$, $P_1 : (x_1, \psi(x_1))$, etc. Estes pontos estão, obviamente, afastando-se da interseção das duas curvas $y_1 = x$ e $y_2 = \psi(x)$, e, ao mesmo tempo, x_k está se afastando de $\bar{x}$. Na Figura 3.8, os pontos P_0, P_1, etc. estão, obviamente, aproximando-se do ponto de interseção das duas curvas $y_1 = x$ e $y_2 = \psi(x)$, e, ao mesmo tempo, x_k está se aproximando de $\bar{x}$.

Assim em a) temos que o processo iterativo é divergente e em b) que o processo iterativo é convergente.

Ordem de Convergência

A ordem de convergência de um método mede a velocidade com que as iterações produzidas por esse método aproximam-se da solução exata. Assim, quanto maior for a ordem de convergência melhor será o método numérico pois mais rapidamente obteremos a solução. Analisaremos aqui a ordem de

convergência do método iterativo linear. Antes porém apresentamos a definição de ordem de convergência de um método numérico.

Definição 3.3 - *Sejam* $\{x_k\}$ *o resultado da aplicação de um método numérico na iteração* k *e* $e_k = x_k - \bar{x}$ *o seu erro . Se existirem um número* $p \geq 1$ *e uma constante* $c > 0$ *tais que:*

$$lim_{k \to \infty} \frac{|e_{k+1}|}{|e_k|^p} = c$$

então p *é chamado de* **ordem de convergência** *desse método.*

Teorema 3.5 - *A ordem de convergência do método iterativo linear é linear, ou seja,* $p = 1$.

Prova: Do teorema 3.4, temos que:

$$x_{k+1} - \bar{x} = \psi'(\xi_k)(x_k - \bar{x})$$

onde ξ_k está entre x_k e $\bar{x}$. Assim,

$$\frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|} \leq |\psi'(\xi_k)| \leq M.$$

Assim a definição 3.3 está satisfeita com $p = 1$ e $c = M$, ou seja a ordem de convergência é $p = 1$. Daí o nome de método Iterativo Linear. Além disso, o erro em qualquer iteração é proporcional ao erro na iteração anterior, sendo que o fator de proporcionalidade é $\psi'(\xi_k)$.

Observações:

a) A convergência do processo iterativo será tanto mais rápida quanto menor for o valor de $\psi'(x)$.

b) Por outro lado, se a declividade $\psi'(x)$ for maior que 1 em valor absoluto, para todo x pertencente a um intervalo numa vizinhança da raiz, vimos que a iteração $x_{k+1} = \psi(x_k), \ k = 0, 1, \ldots$, divergirá.

c) Da definição 3.3 podemos afirmar que para k suficientemente grande temos:

$$\begin{array}{rcl} |e_{k+1}| & \simeq & c\ |e_k|^p \ , \\ |e_k| & \simeq & c\ |e_{k-1}|^p \ . \end{array}$$

Dividindo uma equação pela outra eliminamos a constante c e obtemos:

$$\frac{e_{k+1}}{e_k} \simeq \left(\frac{e_k}{e_{k-1}} \right)^p \ .$$

Assim, uma aproximação para o valor de p pode ser obtido aplicando-se logaritmo em ambos os membros da expressão acima. Fazendo isso segue que:

$$p \ \simeq \ \frac{log\left(\frac{e_{k+1}}{e_k}\right)}{log\left(\frac{e_k}{e_{k-1}}\right)} \tag{3.5}$$

Exemplo 3.8 - *Com os valores obtidos no exemplo 3.7, verifique que o método iterativo linear realmente possui ordem de convergência* $p = 1$.

Solução: Do resultado do exemplo 3.7, fazendo os cálculos para os valores de $|x_{k+1} - \bar{x}|$ e usando (3.5), obtemos a tabela:

k	$x_{k+1} = \sqrt{2 + x_k}$	$e_k = \|x_k - \bar{x}\|$	p
0	2.5	0.5	
1	2.1213203	0.1213203	
2	2.0301035	0.0301035	0.984
3	2.0075118	0.0075118	0.996
4	2.0018771	0.0018771	0.999
5	2.0004692	0.0004692	0.999
6	2.0001173	0.0001173	0.999
7	2.0000293	0.0000293	1.001

Pela tabela vemos que a medida que k aumenta, o valor de $p \to 1$, mostrando que realmente a ordem de convergência do método iterativo linear é 1.

Assim podemos dizer que a importância do método iterativo linear está mais nos conceitos que são introduzidos em seu estudo que em sua eficiência computacional. Além disso, tem a desvantagem de que é preciso testar se $|\psi'(x)| < 1$ no intervalo que contém a raiz, se desejamos ter garantia de convergência.

Exercícios

3.3 - *Justifique que a equação:* $f(x) = 4\ x - e^x = 0$ *possui uma raiz no intervalo* $(0, 1)$ *e outra no intervalo* $(2, 3)$.

3.4 - *Considere a equação* $f(x) = 2\ x^2 - 5\ x + 2 = 0$, *cujas raízes são:* $x_1 = 0.5$ *e* $x_2 = 2.0$. *Considere ainda os processos iterativos:*

a) $x_{k+1} = \dfrac{2\ x_k^2 + 2}{5}$,

b) $x_{k+1} = \sqrt{\dfrac{5\ x_k}{2} - 1}$.

Qual dos dois processos você utilizaria para obter a raiz x_1*? Por que?*

3.5 - *Considere as seguintes funções:*

a) $\psi_1(x) = 2\ x - 1$,

b) $\psi_2(x) = x^2 - 2\ x + 2$,

c) $\psi_3(x) = x^2 - 3\ x + 3$.

Verifique que 1 é raiz de todas estas funções. Qual delas você escolheria para obter a raiz 1, utilizando o processo iterativo $x_{k+1} = \psi(x_k)$*? Com a sua escolha, exiba a sequência gerada a partir da condição inicial* $x_0 = 1.2$.

3.6 - *Deseja-se obter a raiz positiva da equação:* $b\ x^2 + x - a = 0, a > 0,\ b > 0$, *através do processo iterativo definido por:* $x_{k+1} = a - b\ x_k^2$ *.Qual a condição que devemos impor para a e b para que haja convergência? Por que?*

3.7 - *A equação:* $x^2 - a = 0$ *possui uma raiz* $\bar{x} = \sqrt{a}$. *Explicar algébrica e geometricamente por que a sequência* $\{x_k\}$, *obtida através do processo iterativo definido por:* $x_{k+1} = \dfrac{a}{x_k}$ *, não converge para* $\sqrt{a}$ *qualquer que seja o valor de* x_0.

3.3 Método de Newton

O método de Newton é uma das técnicas mais populares para se determinar raízes de equações não lineares. Existem várias maneiras de deduzir o método de Newton, a que apresentaremos aqui é baseada no método de iteração linear. Assim, para descrever tal método, consideremos a equação (3.2), isto é:

$$\psi(x) = x + A(x)f(x) \ , \quad \text{com} \quad f'(x) \neq 0 \ , \tag{3.6}$$

onde a função $A(x)$ deve ser escolhida de tal forma que $A(\bar{x}) \neq 0$.

Vimos pelo Teorema 3.4 que temos garantia de convergência se $max|\psi'(x)| < 1$ para $x \in I$. Assim se escolhermos $A(x)$ tal que $\psi'(\bar{x}) = 0$, teremos que para $x \in I$, (I suficientemente pequeno), $\max|\psi'(x)| < 1$, garantindo então a convergência do método.

Derivando (3.6) em relação a x, obtemos:

$$\psi'(x) \ = \ 1 \ + \ A'(x)f(x) \ + \ A(x)f'(x).$$

Fazendo $x = \bar{x}$, segue que:

$$\psi'(\bar{x}) \ = \ 1 \ + \ A(\bar{x})f'(\bar{x}) \ , \quad \text{pois} \quad f(\bar{x}) = 0 \ ,$$

e colocando:

$$\psi'(\bar{x}) \ = \ 0, \quad \text{teremos} \quad A(\bar{x}) \ = \ - \ \frac{1}{f'(\bar{x})} \neq 0 \quad \text{desde que} \quad f'(\bar{x}) \neq 0 \ .$$

Tomando então: $A(x) \ = \ - \ \frac{1}{f'(x)}$, obtemos $\psi(x) \ = \ x \ - \ \frac{f(x)}{f'(x)}$ e o processo iterativo então definido por:

$$x_{k+1} \ = \ x_k \ - \ \frac{f\,(x_k)}{f'\,(x_k)} \tag{3.7}$$

é chamado **Método de Newton**, que converge sempre que $|x_0 - \bar{x}|$ for suficientemente pequeno.

Uma interpretação geométrica do método de Newton é dada na Figura 3.9.

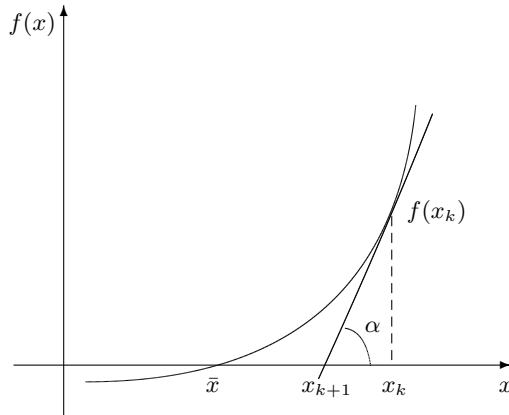



Figura 3.9

Dado x_k, o valor x_{k+1} pode ser obtido graficamente traçando-se pelo ponto $(x_k, f(x_k))$ a tangente à curva $y = f(x)$. O ponto de interseção da tangente com o eixo dos x determina x_{k+1}.

De fato, pela lei da tangente:

$$f'(x_k) \; = \; tg \; \alpha \quad = \quad \frac{f(x_k)}{x_k - x_{k+1}}$$

$$\Rightarrow \; x_k \; - \; x_{k+1} \; = \; \frac{f(x_k)}{f'(x_k)} \quad \Rightarrow \quad x_{k+1} \; = \; x_k \; - \; \frac{f(x_k)}{f'(x_k)} \; .$$

Devido à sua interpretação geométrica o método de Newton é também chamado de **Método das Tangentes**.

Exemplo 3.9 - *Determinar, usando o método de Newton, a menor raiz positiva da equação:*

$$4 \; cos \; x \; - \; e^x \; = \; 0 \; ,$$

com erro inferior a 10^{-2}.

Solução: O processo mais simples e eficaz para se obter um valor inicial é o método gráfico. Com esse objetivo dividimos a equação inicial $f(x) \; = \; 0$ em outras duas equações mais simples, que chamaremos de y_1 e y_2. Note que o rearranjo para obter essas duas equações deve apenas levar em consideração a igualdade $f(x) = 0$.

Tomando: $y_1 \; = \; 4 \; cos \; x$, $y_2 \; = \; e^x$, observe que poderíamos ter tomado $y_1 \; = \; cos \; x$ e $y_2 \; = \; \frac{e^x}{4}$, e colocando as duas funções no mesmo gráfico, obtemos a Figura 3.10.

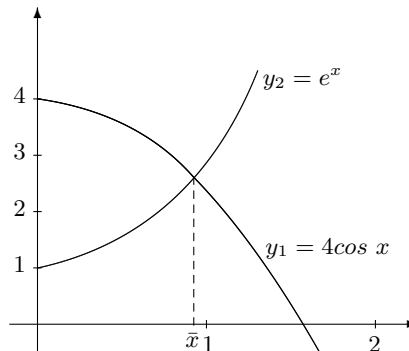



Figura 3.10

Como já dissemos anteriormente, o ponto de interseção das duas curvas é a solução $\bar{x}$ procurada. Analisando a Figura 3.10, vemos que $\bar{x}$ está nas vizinhanças do ponto 1.0 e portanto vamos tomar $x_0 = 1.0$. Por outro lado, da equação original, obtemos:

$$f(x_k) \; = \; 4 \; cos \; x_k \; - \; e^{x_k} \; ;$$

$$f'(x_k) \; = \; -4 \; sen \; x_k \; - \; e^{x_k} \; .$$

Para efetuar os cálculos seguintes, observe se sua calculadora está em radianos, pois a função dada envolve operações trigonométricas. Além disso, como queremos o resultado com erro inferior a 10^{-2} basta efetuar os cálculos com três casas decimais. Assim:

$$\begin{aligned} f(x_0) \; &= \; f(1.0) \; = \; 4 \; cos \; (1.0) - e^{1.0} \\ &= \; 4 \; \times \; (0.540) - 2.718 \; = \; -0.557 \; ; \\ f'(x_0) \; &= \; f'(1.0) \; = \; -4 \; sen \; (1.0) - e^{1.0} \\ &= \; -4 \; \times \; (0.841) - 2.718 \; = \; -6.084 \; ; \end{aligned}$$

Usando (3.7), obtemos:

$$x_1 \;=\; 1.0 \;-\; \frac{f(1.0)}{f'(1.0)} \Rightarrow \; x_1 \;=\; 1.0 \;-\; \frac{(-0.557)}{(-6.048)} \;\Rightarrow\; x_1 \;=\; 0.908 \; .$$

Calculando o erro relativo, temos:

$$\left| \frac{x_1 - x_0}{x_1} \right| \;\simeq\; 0.101 \; .$$

que é maior que 10^{-2}. Devemos fazer uma nova iteração, para tanto calculemos:

$$\begin{aligned} f(x_1) &= f(0.908) \;=\; 4\; cos\; (0.908) - e^{0.908} \\ &= 4 \;\times\; (0.615) - 2.479 \;=\; -0.019 \;, \\ f'(x_1) &= f'(0.908) \;=\; -4\; sen\; (0.908) - e^{0.908} \\ &= -4 \;\times\; (0.788) - 2.479 \;=\; -5.631 \; . \end{aligned}$$

Novamente, usando (3.7), obtemos:

$$x_2 \;=\; 0.908 \;-\; \frac{f(0.908)}{f'(0.908)} \;\Rightarrow\; x_2 \;=\; 0.908 \;-\; \frac{(-0.019)}{(-5.631)} \;\Rightarrow\; x_2 \;=\; 0.905 \; .$$

Calculando o erro relativo, segue que:

$$\left| \frac{x_2 - x_1}{x_2} \right| \;\simeq\; 0.0033 \;,$$

ou seja a aproximação $x_2 = 0.905$ possui duas casas decimais corretas. De fato, a solução exata é 0.9047882. Logo, a menor raiz positiva da equação $4\; cos\; x - e^x = 0$, com $\epsilon < 0.01$, é $\bar{x} = 0.905$. Observe, da Figura 3.10, que a raiz encontrada é a única raiz positiva da equação dada.

Ordem de Convergência

Analisemos agora a convergência do método de Newton.

Teorema 3.6 - *Se f, f', f'' são contínuas em I cujo centro $\bar{x}$ é solução de $f(x) = 0$ e se $f'(\bar{x}) \neq 0$ então a ordem de convergência do método de Newton é quadrática, ou seja, $p = 2$.*

Prova: Subtraindo a equação $\bar{x} = \psi(\bar{x})$ de (3.7), obtemos:

$$x_{k+1} \;-\; \bar{x} \;=\; \psi(x_k) \;-\; \psi(\bar{x}) \text{ onde } \psi(x) \;=\; x \;-\; \frac{f(x)}{f'(x)} \; .$$

Desenvolvendo $\psi(x_k)$ em série de Taylor em torno do ponto $\bar{x}$, obtemos:

$$x_{k+1} \;-\; \bar{x} \;=\; \psi(\bar{x}) \;+\; (x_k - \bar{x})\; \psi'(\bar{x}) \;+\; \frac{(x_k - \bar{x})^2}{2!} \psi''(\xi_k) \;-\; \psi(\bar{x}) \; .$$

Agora, no método de Newton, $\psi'(\bar{x}) = 0$, (lembre-se que esta condição foi imposta para a determinação de $A(x)$), e portanto:

$$\frac{|x_{k+1} - \bar{x}|}{|x_k - \bar{x}|^2} \;=\; \frac{\psi''(\xi_k)}{2!} \;\leq\; C.$$

Portanto a ordem de convergência é $p = 2$.

Assim a vantagem do método de Newton é que sua convergência é quadrática, isto significa que a quantidade de dígitos significativos corretos duplica à medida que os valores da sequência se aproxima de $\bar{x}$. Note que essa correção não acontece em relação às primeiras iterações realizadas. A desvantagem do método de Newton está no fato de termos que calcular a derivada da função e em cada iteração calcular o seu valor numérico, o que pode ser muito caro computacionalmente. Além disso a função pode ser não diferenciável em alguns pontos do domínio.

Exercícios

3.8 - *Considere a equação dada no exemplo 3.9. Obtenha a raiz positiva com quatro casas decimais corretas. Usando (3.5) confirme que a ordem de convergência do método de Newton é quadrática, isto é,* $p = 2$.

3.9 - *Usando o método de Newton, com erro inferior a* 10^{-2}*, determinar uma raiz das seguintes equações:*

a) $2\ x = tg\ x,$

b) $5\ x^3 + x^2 - 12\ x + 4 = 0,$

c) $sen\ x - e^x = 0,$

d) $x^4 - 8 = 0.$

3.10 - *Considere a fórmula para determinar a raiz cúbica de* Q*:*

$$x_{k+1} \ = \ \frac{1}{3}\left[2x_k \ + \ \frac{Q}{x_k^2}\right] \ , \quad k \ = \ 0, 1, \ldots \ .$$

a) *Mostre que a fórmula acima é um caso especial de iteração de Newton.*

b) *Usando a fórmula dada no item a) calcule* $\sqrt[3]{4}$ *com precisão de* 10^{-2}*, determinando o valor inicial através de gráfico.*

3.11 - *Usando o método de Newton, determine o valor de* π *com 3 algarismos significativos corretos. Use como valor inicial* $x_0 = 3$.

3.4 Método das Secantes

Como foi observado anteriormente, uma séria desvantagem do método de Newton é a necessidade de se obter $f'(x)$, bem como calcular seu valor numérico, a cada passo. Há várias maneiras de modificar o método de Newton a fim de eliminar essa desvantagem. Uma modificação consiste em substituir a derivada $f'(x_k)$ pelo quociente das diferenças:

$$f'(x_k) \ \cong \ \frac{f(x_k) \ - \ f(x_{k-1})}{x_k \ - \ x_{k-1}} \ , \tag{3.8}$$

onde x_k, x_{k-1} são duas aproximações quaisquer para a raiz $\bar{x}$. Note que $f'(x_k)$ é o limite da relação (3.8) para $x_{k-1} \ \rightarrow \ x_k$.

O método de Newton, quando modificado desta maneira, é conhecido como **Método das Secantes**. Substituindo (3.8) em (3.7) obtemos:

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)}{\dfrac{f(x_k) - f(x_{k-1})}{(x_k - x_{k-1})}} \\ &= x_k - \frac{(x_k - x_{k-1}) f(x_k)}{f(x_k) - f(x_{k-1})} \end{aligned}$$

Assim, colocando o segundo membro sobre o mesmo denominador, obtemos uma expressão mais simples para o método das secantes:

$$x_{k+1} = \frac{x_{k-1} f(x_k) - x_k f(x_{k-1})}{f(x_k) - f(x_{k-1})} . \quad (3.9)$$

Observe que devem estar disponíveis duas aproximações iniciais, antes que (3.9) possa ser usada. Na Figura 3.11, ilustramos graficamente como uma nova aproximação, pode ser obtida de duas anteriores.

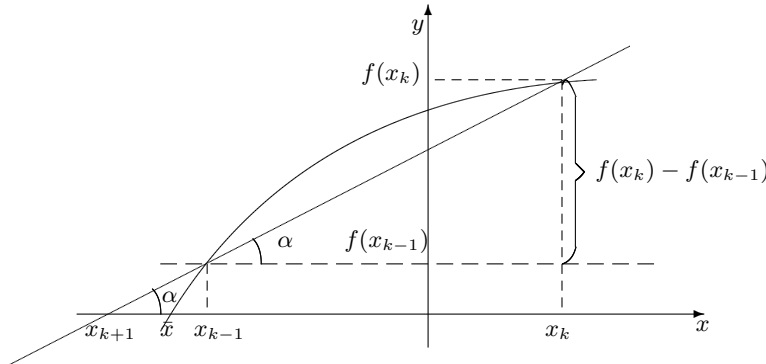



Figura 3.11

Pela Figura 3.11 vemos que, geometricamente, o método das secantes consiste em considerar, como aproximação seguinte, a interseção da corda que une os pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$ com o eixo dos x. Tomando:

$$\begin{aligned} x_{k+1} &= x_k - \frac{f(x_k)\ (x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} \\ \Rightarrow \frac{x_{k+1} - x_k}{f(x_k)} &= \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \\ \Rightarrow \frac{f(x_k)}{x_{k+1} - x_k} &= \frac{f(x_k) - f(x_{k-1})}{x_{k-1} - x_k} = tg\ \alpha. \end{aligned}$$

Exemplo 3.10 - *Determinar a raiz positiva da equação:*

$$\sqrt{x} - 5\ e^{-x} = 0 ,$$

pelo método das secantes, com erro inferior a 10^{-2}.

Solução: Novamente, para obtermos os valores iniciais x_0 e x_1 necessários para iniciar o processo iterativo, dividimos a equação original $f(x) = 0$ em outras duas y_1 e y_2, com $y_1 = \sqrt{x}$ e $y_2 = 5e^{-x}$, que colocadas no mesmo gráfico, produz a Figura 3.12:

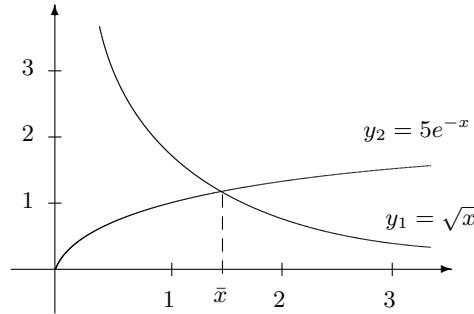



Figura 3.12

O ponto de interseção das duas curvas é a solução $\bar{x}$ procurada. Analisando a Figura 3.12, vemos que $\bar{x}$ está nas vizinhanças do ponto 1.4. Assim, tomando $x_0 = 1.4$ e $x_1 = 1.5$, obtemos:

$$\begin{aligned} f(x_0) &= f(1.4) = \sqrt{1.4} - 5\ e^{-1.4} = 1.183 - 5 \times 0.247 = -0.052\ , \\ f(x_1) &= f(1.5) = \sqrt{1.5} - 5\ e^{-1.5} = 1.225 - 5 \times 0.223 = 0.110\ . \end{aligned}$$

Portanto, usando (3.9), obtemos que:

$$\begin{aligned} & x_2 = \frac{1.4\ f(1.5) - 1.5\ f(1.4)}{f(1.5) - f(1.4)}. \\ \Rightarrow\ & x_2 = \frac{1.4\ (0.110) - 1.5\ (-0.052)}{0.110 - (-0.052)} \\ \Rightarrow\ & x_2 = 1.432\ . \end{aligned}$$

Calculando o erro relativo:

$$\left| \frac{x_2 - x_1}{x_2} \right| \simeq 0.047$$

observamos que este é maior que 10^{-2}. Devemos portanto fazer mais uma iteração. Calculemos então:

$$\begin{aligned} f(x_2) &= f(1.432) = \sqrt{1.432} - 5\ e^{-1.432} \\ &= 1.197 - 5 \times 0.239 = 0.002. \end{aligned}$$

Novamente, usando (3.9), obtemos:

$$\begin{aligned} & x_3 = \frac{1.5\ f(1.432) - 1.432\ f(1.5)}{f(1.432) - f(1.5)}. \\ \Rightarrow\ & x_3 = \frac{1.5\ (0.002) - 1.432\ (0.110)}{0.002 - (0.110)} \\ \Rightarrow\ & x_3 = 1.431\ . \end{aligned}$$

Fazendo:

$$\left| \frac{x_3 - x_2}{x_3} \right| \simeq 0.0007 < 10^{-2}.$$

Logo, a raiz positiva da equação $\sqrt{x} - 5\ e^{-x} = 0$, com $\epsilon < 10^{-2}$, é $\bar{x} = 1.431$.

Ordem de Convergência

Daremos aqui a ordem de convergência do método das secantes.

Teorema 3.7 - *A ordem de convergência do método das secantes é* $p = (1+\sqrt{5})/2 \simeq 1.618$.

Prova: A prova deste teorema pode ser encontrada em [Ostrowski,1966].

Observe que apesar da ordem de convergência do método das secantes ser inferior à do método de Newton, ele fornece uma alternativa viável, desde que requer somente um cálculo da função f por passo, enquanto que dois cálculos ($f(x_k)$ e $f'(x_k)$) são necessários para o método de Newton.

Exercícios

3.12 - *Considere a equação dada no exemplo 3.10. Obtenha a raiz positiva com quatro casas decimais corretas. Usando (3.5) confirme que a ordem de convergência do método das secantes é* $p \simeq 1.618$.

3.13 - *Determinar, pelo método das secantes, uma raiz de cada uma das equações:*

a) $x = 2.7 \; ln \; x,$

b) $log \; x \; - \; cos \; x = 0,$

c) $e^{-x} \; - \; log \; x = 0.$

3.14 - *A equação* $x = tgx$ *tem uma raiz entre* $\frac{\pi}{2}$ *e* $\frac{3\pi}{2}$. *Determiná-la pelo método das secantes com erro inferior a* 10^{-3}.

3.5 Método Regula Falsi

O método **Regula Falsi**, é uma variação do método das secantes. Ele consiste em tomar duas aproximações iniciais x_0 e x_1 tais que $f(x_0)$ e $f(x_1)$ tenham sinais opostos, isto é:

$$f(x_0) \; \times \; f(x_1) \; < \; 0.$$

Uma nova aproximação é determinada usando o método das secantes, ou seja:

$$x_2 \; = \; \frac{x_0 \; f(x_1) \; - \; x_1 \; f(x_0)}{f(x_1) \; - \; f(x_0)}.$$

Se

$$\left|\frac{x_2 - x_0}{x_2}\right| \; < \; \epsilon \quad \text{ou} \left|\frac{x_2 - x_1}{x_2}\right| \; < \; \epsilon \; ,$$

para um ϵ pré-fixado, então x_2 é a raiz procurada. Caso contrário, calculamos $f(x_2)$ e escolhemos entre x_0 e x_1 aquele cuja f tenha sinal oposto ao da $f(x_2)$. Com x_2 e esse ponto calculamos x_3 usando a fórmula das secantes, isto é, usando (3.9) e assim sucessivamente. O processo iterativo deve ser continuado até que se obtenha a raiz com a precisão pré-fixada.

Uma interpretação geométrica do método regula falsi é dada na Figura 3.13. Observe que, na Figura 3.13, x_{k+1} é o ponto de interseção da corda que une os pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_k, f(x_k))$ com o eixo dos x. Nesse caso o novo intervalo contendo a raiz será (x_{k-1}, x_{k+1}). A aproximação x_{k+2} será o ponto de interseção da corda que une os pontos $(x_{k-1}, f(x_{k-1}))$ e $(x_{k+1}, f(x_{k+1}))$ com o eixo dos x. Observe ainda que a aplicação do método regula falsi sempre mantém a raiz procurada entre as aproximações mais recentes.

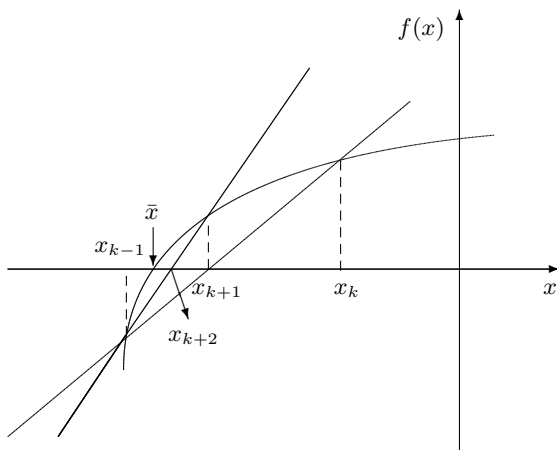



Figura 3.13

Exemplo 3.11 - *Determinar a menor raiz positiva da equação:*

$$x \; - \; cos \; x = 0 \; ,$$

pelo método regula falsi, com erro inferior a 10^{-3}.

Solução: Novamente, para obtermos os valores iniciais x_0 e x_1 necessários para iniciar o processo iterativo, dividimos a equação original $f(x) = 0$ em $y_1 = x$ e $y_2 = cos \; x$ que colocadas no mesmo gráfico, produz a Figura 3.14:

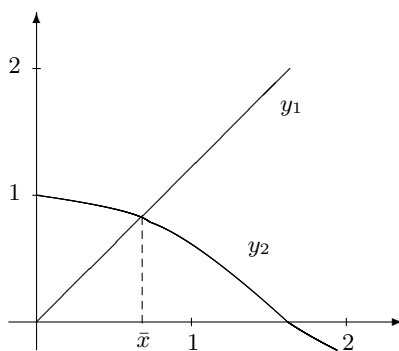



Figura 3.14

O ponto de interseção das duas curvas é a solução $\bar{x}$ procurada. Analisando a Figura 3.14, vemos que $\bar{x}$ está nas vizinhanças do ponto 0.7. Assim, tomando $x_0 = 0.7$ e $x_1 = 0.8$, obtemos:

$$\begin{aligned} f\,(x_0) &= f(0.7) = 0.7 - cos\; 0.7 = 0.7 - 0.7648 = -0.0648 \; , \\ f\,(x_1) &= f(0.8) = 0.8 - cos\; 0.8 = 0.8 - 0.6967 = 0.1033 \; , \end{aligned}$$

e portanto: $f(x_0) \times f(x_1) < 0$. Portanto, usando (3.9), obtemos que:

$$\begin{aligned} x_2 &= \frac{0.7\, f(0.8) - 0.8\, f(0.7)}{f(0.8) - f(0.7)}. \\ \Rightarrow \quad x_2 &= \frac{0.7\,(0.1033) - 0.8\,(-0.0648)}{0.1033 - (-0.0648)} \\ \Rightarrow \quad x_2 &= 0.7383. \end{aligned}$$

Fazendo:

$$\left| \frac{x_2 - x_0}{x_2} \right| \simeq 0.052 \text{ e } \left| \frac{x_2 - x_1}{x_2} \right| \simeq 0.084$$

obtemos que ambos são maiores que 10^{-3}. Calculando:

$$\begin{aligned} f(x_2) &= f(0.7383) = 0.7383 - \cos 0.7383 \\ &= 0.7383 - 0.7396 = -0.0013. \end{aligned}$$

vemos que $f(x_2) \times f(x_1) < 0$, e portanto a raiz está entre x_1 e x_2. Assim, usando novamente (3.9), segue que:

$$\begin{aligned} x_3 &= \frac{0.8\, f(0.7383) - 0.7383\, f(0.8)}{f(0.7383) - f(0.8)}. \\ \Rightarrow \quad x_3 &= \frac{0.8\,(-0.0013) - 0.7383\,(0.1033)}{-0.0013 - (0.1033)} \\ \Rightarrow \quad x_3 &= 0.7390. \end{aligned}$$

Calculando o erro relativo:

$$\left| \frac{x_3 - x_2}{x_3} \right| \simeq 0.00095 \,,$$

vemos que este é menor 10^{-3}. Assim, a menor raiz positiva, (observe pela Figura 3.14, que a raiz positiva é única) , da equação $x - \cos x = 0$, com $\epsilon < 10^{-3}$, é $\bar{x} = 0.7390$.

Ordem de Convergência

A ordem de convergência do método regula falsi é semelhante ao do método das secantes uma vez que o procedimento para o cálculo das aproximações são os mesmos em ambos os casos. Assim a ordem de convergência do método regula falsi também é $p = (1 + \sqrt{5})/2 \simeq 1.618$.

Exercícios

3.15 - *Considere a equação dada no exemplo 3.11. Obtenha a raiz positiva com cinco casas decimais corretas. Usando (3.5) confirme que a ordem de convergência do método regula falsi é* $p \simeq 1.618$.

3.16 - *Determinar uma raiz de cada uma das equações:*

a) $sen\, x - x\, e^x = 0$,

b) $\cos x = e^x$,

usando o método regula falsi.

3.17 - *A equação:* $x - 2\, sen\, x = 0$ *possui uma raiz no intervalo* $[1.8, 2.0]$. *Determiná-la pelo método regula falsi, com duas casas decimais corretas.*

3.6 Sistemas de Equações não Lineares

Nesta seção consideramos o problema da determinação de raízes de equações não lineares simultâneas da forma:

$$\begin{cases} f_1(x_1, x_2, \ldots, x_m) = 0 \\ f_2(x_1, x_2, \ldots, x_m) = 0 \\ \vdots \\ f_m(x_1, x_2, \ldots, x_m) = 0 \end{cases}$$

onde cada $f_i, i = 1, 2, \ldots, m$ é uma função real de m variáveis reais. Embora esse tópico seja de considerável importância, daremos aqui apenas uma breve introdução. Para maiores detalhes os interessados podem consultar, por exemplo, [Ortega, 70].

Assim, para efeito de simplicidade, e sem perda de generalidade, consideraremos apenas o caso de duas equações a duas incógnitas, isto é, sistemas não lineares da forma:

$$\begin{cases} f(x, y) = 0 \\ g(x, y) = 0 \end{cases} \tag{3.10}$$

Geometricamente, as raízes deste sistema são os pontos do plano (x, y), onde as curvas definidas por f e g se interceptam.

3.6.1 Iteração Linear

A resolução de sistemas não lineares através do método de iteração linear é muito semelhante ao método iterativo linear estudado anteriormente. Assim, um primeiro passo ao se aplicar iteração linear é reescrever o sistema (3.10) na forma:

$$\begin{cases} x = F(x, y) \\ y = G(x, y). \end{cases} \tag{3.11}$$

de forma que qualquer solução de (3.11) seja, também, solução de (3.10).

Sejam $(\bar{x}, \bar{y})$ uma solução de (3.10) e (x_0, y_0) uma aproximação para $(\bar{x}, \bar{y})$. Obtemos as aproximações sucessivas (x_k, y_k) para a solução desejada $(\bar{x}, \bar{y})$, usando o processo iterativo definido por:

$$\begin{cases} x_{k+1} = F(x_k, y_k) \\ y_{k+1} = G(x_k, y_k). \end{cases} \tag{3.12}$$

Esse processo é chamado **Método Iterativo Linear para Sistemas não Lineares**.

O processo (3.12) convergirá sob as seguintes condições suficientes (mas não necessárias):

a) F, G e suas derivadas parciais de primeira ordem sejam contínuas numa vizinhança V da raiz $(\bar{x}, \bar{y})$.

b) As seguintes desigualdades sejam satisfeitas:

$$\begin{aligned} |F_x| + |F_y| &\leq k_1 < 1 \\ |G_x| + |G_y| &\leq k_2 < 1, \end{aligned}$$

para todo ponto (x, y) pertencente à uma vizinhança V de $(\bar{x}, \bar{y})$, onde: $F_x = \dfrac{\partial F}{\partial x}, F_y = \dfrac{\partial F}{\partial y}$, etc.

c) A aproximação inicial (x_0, y_0) pertença a vizinhança V de $(\bar{x}, \bar{y})$.

Para obtermos uma solução com uma determinada precisão ϵ devemos, durante o processo iterativo, calcular o erro relativo para todas as componentes do vetor solução.

Exemplo 3.12 - *Considere o seguinte sistema não linear:*

$$\begin{cases} f(x,y) &= 0.2\ x^2 + 0.2\ x\ y - x + 0.6 = 0 \\ g(x,y) &= 0.4\ x + 0.1 x\ y^2 - y + 0.5 = 0 \end{cases}$$

a) *Verifique que reescrevendo o sistema dado na forma:*

$$\begin{cases} x &= 0.2\ x^2 + 0.2\ x\ y + 0.6 \ = \ F(x,y) \\ y &= 0.4\ x + 0.1\ x\ y^2 + 0.5 \ = \ G(x,y) \end{cases}$$

as condições suficientes para garantir a convergência são satisfeitas.

b) *Aplique o método iterativo linear para resolver o sistema dado.*

Solução: Uma solução desse sistema, facilmente comprovável, é o ponto:$(\bar{x},\bar{y}) = (1,1)$. É claro que não conhecemos, a priori, a solução do sistema , mas este é apenas um exemplo para ilustrar a verificação das condições suficientes de convergência , bem como a aplicação do método iterativo linear. Mais adiante mostraremos como determinar os valores iniciais.

Para verificar as condições suficientes, calculemos inicialmente, as derivadas parciais de F e G. Assim:

$$\begin{array}{lll} F_x = 0.4\ x + 0.2\ y\ , & & F_y = 0.2\ x\ , \\ G_x = 0.4 + 0.1\ y^2\ , & & G_y = 0.2\ x\ y\ , \end{array}$$

Se escolhermos, por exemplo, $(x_0, y_0) = (0.9, 1.1)$, vemos que F, G e suas derivadas parciais são contínuas em (x_0, y_0). Além disso, as desigualdades que figuram nas condições para convergência, são satisfeitas, pois temos:

$$\begin{array}{l} |F_x| + |F_y| = |(0.4)(0.9)| + |(0.2)(1.1)| + |(0.2)(0.9)| = 0.76 < 1 \\ |G_x| + |G_y| = |(0.4) + (0.1)(1.1)^2| + |(0.2)(0.9)(1.1)| = 0.719 < 1. \end{array}$$

E é claro que (x_0, y_0) está na vizinhança de $(\bar{x},\bar{y})$. Tomando então $(x_0, y_0) = (0.9, 1.1)$ e usando o processo iterativo definido por(3.12), obtemos:

$$\begin{array}{lll}
x_1 & = & F(x_0, y_0) = (0.2)(0.9)^2 + (0.2)(0.5)(1.1) + 0.6 \\
& \Rightarrow & x_1 = 0.96 \\
y_1 & = & G(x_0, y_0) = (0.4)(0.9) + (0.1)(0.9)(1.1)^2 + 0.5 \\
& \Rightarrow & y_1 = 0.9689 \\
x_2 & = & F(x_1, y_1) = (0.2)(0.96)^2 + (0.2)(0.96)(0.9689) + 0.6 \\
& \Rightarrow & x_2 = 0.9703 \\
y_2 & = & G(x_1, y_1) = (0.4)(0.96) + (0.1)(0.96)(0.0.9689)^2 + 0.5 \\
& \Rightarrow & y_2 = 0.9791 \\
x_3 & = & F(x_2, y_2) = (0.2)(0.9703)^2 + (0.2)(0.9703)(0.9791) + 0.6 \\
& \Rightarrow & x_3 = 0.9773 \\
y_3 & = & G(x_2, y_2) = (0.4)(0.9703) + (0.1)(0.9703)(0.9791)^2 + 0.5 \\
& \Rightarrow & y_3 = 0.9802 \\
\vdots & &
\end{array}$$

É claro que a sequência (x_k, y_k) está convergindo para $(1,1)$. Além disso, podemos dizer que a solução $(\bar{x},\bar{y})$, com erro relativo inferior a 10^{-2}, é $(0.9773, 0.9802)$, desde que $\dfrac{|x_3 - x_2|}{x_3} \simeq 0.007$ e $\dfrac{|y_3 - y_2|}{y_3} \simeq 0.001$. Observe ainda que mesmo se uma das componentes estiver com a precisão desejada, mas a outra não, o processo deve ser continuado até que todas estejam com a precisão pré-fixada.

Exercícios

3.18 - *Usando o método iterativo linear determinar a solução de:*

$$\begin{cases} x = 0.7 \; sen \; x + 0.2 \; cos \; y \\ y = 0.7 \; cos \; x - 0.2 \; sen \; y \end{cases}$$

próxima a (0.5,0.5).

3.19 - *O sistema não linear:*

$$\begin{cases} x^2 + x\,y^2 = 2 \\ x\,y - 3\,x\,y^3 = -4 \end{cases}$$

possui uma raiz próxima a (0.8,1.2). Usando o método iterativo linear determine essa raiz com precisão de 10^{-1}.

3.6.2 Método de Newton

Para adaptar o método de Newton a sistemas não lineares, procedemos como se segue:

Seja (x_0, y_0) uma aproximação para a solução $(\bar{x}, \bar{y})$ de (3.10). Admitindo que f e g sejam suficientemente diferenciáveis, expandimos $f(x, y)$ e $g(x, y)$, usando série de Taylor para funções de duas variáveis, em torno de (x_0, y_0). Assim:

$$\begin{cases} f(x,y) = f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) + \ldots \\ g(x,y) = g(x_0,y_0) + g_x(x_0,y_0)(x-x_0) + g_y(x_0,y_0)(y-y_0) + \ldots \end{cases}$$

Admitindo que (x_0, y_0) esteja suficientemente próximo da solução $(\bar{x}, \bar{y})$ a ponto de poderem ser abandonados os termos de mais alta ordem, podemos determinar uma nova aproximação para a raiz $(\bar{x}, \bar{y})$ fazendo $f(x, y) = g(x, y) = 0$. Obtemos, então, o sistema:

$$\begin{cases} f_x(x-x_0) + f_y(y-y_0) = -f \\ g_x(x-x_0) + g_y(y-y_0) = -g \end{cases} \tag{3.13}$$

onde está entendido que todas as funções e derivadas parciais em (3.13) devem ser calculadas em (x_0, y_0). Observe que (3.13) é agora um sistema linear. Além disso, se não tivessemos desprezado os termos de mais alta ordem no desenvolvimento de Taylor , então (x, y) seria a solução exata do sistema não linear. Entretanto, a resolução de (3.13) fornecerá uma solução que chamaremos de (x_1, y_1). Devemos, então, esperar que (x_1, y_1) esteja mais próxima de $(\bar{x}, \bar{y})$ do que (x_0, y_0).

Resolvendo (3.13) pela regra de Cramer obtemos:

$$x_1 - x_0 = \frac{\begin{vmatrix} -f & f_y \\ -g & g_y \end{vmatrix}}{\begin{vmatrix} f_x & f_y \\ g_x & g_y \end{vmatrix}} = \left[\frac{-fg_y + gf_y}{J(f,g)}\right]_{(x_0,y_0)}$$

$$y_1 - y_0 = \frac{\begin{vmatrix} f_x & -f \\ g_x & -g \end{vmatrix}}{\begin{vmatrix} f_x & f_y \\ g_x & g_y \end{vmatrix}} = \left[\frac{-gf_x + fg_x}{J(f,g)}\right]_{(x_0,y_0)}$$

onde $J(f, g) = f_x g_y - f_y g_x \neq 0$ em (x_0, y_0). A função $J(f, g)$ é denominada de **Jacobiano** das funções f e g. A solução (x_1, y_1) desse sistema fornece, agora, uma nova aproximanção para $(\bar{x}, \bar{y})$. A repetição

desse processo conduz ao **Método de Newton para Sistemas não Lineares**.

Assim, o método de Newton para sistemas não lineares é definido por:

$$\begin{cases} x_{k+1} = x_k - \left[\dfrac{f g_y - g f_y}{J(f,g)}\right]_{(x_k,\, y_k)} \\ y_{k+1} = y_k - \left[\dfrac{g f_x - f g_x}{J(f,g)}\right]_{(x_k,\, y_k)} \end{cases} \tag{3.14}$$

com $J(f,g) = f_x g_y - f_y g_x$.

Observações:

1) Quando essa iteração converge, a convergência é quadrática.

2) O método de Newton converge sob as seguintes condições suficientes:

- **a)** f, g e suas derivadas parciais até segunda ordem sejam contínuas e limitadas numa vizinhança V contendo $(\bar{x}, \bar{y})$.
- **b)** O Jacobiano $J(f,g)$ não se anula em V.
- **c)** A aproximação inicial (x_0, y_0) seja escolhida suficientemente próxima da raiz $(\bar{x}, \bar{y})$.

3) O método de Newton pode ser, obviamente, aplicado a um sistema de n equações a n incógnitas. Em cada etapa da iteração teremos, então, que calcular n^2 funções derivadas parciais e n funções. Isso representa um considerável custo computacional. Novamente, a menos que seja disponível uma informação, a priori, a respeito da localização da raiz desejada, há, claramente, a possibilidade da iteração não convergir ou que ela convirja para uma outra raiz. A solução de um sistema de n equações, sendo n um valor elevado, torna-se muito difícil mesmo com o uso de computadores.

Exemplo 3.13 - *Determinar uma raiz do sistema:*

$$\begin{cases} x^2 + y^2 = 2 \\ x^2 - y^2 = 1 \end{cases}$$

com precisão de 10^{-3}*, usando o método de Newton.*

Solução: Temos: $f(x,y) = x^2 + y^2 - 2 = 0$ e $g(x,y) = x^2 - y^2 - 1 = 0$. Para obter o valor inicial (x_0, y_0), traçamos no mesmo gráfico as duas equações dadas. Para o sistema dado, obtemos a Figura 3.15:

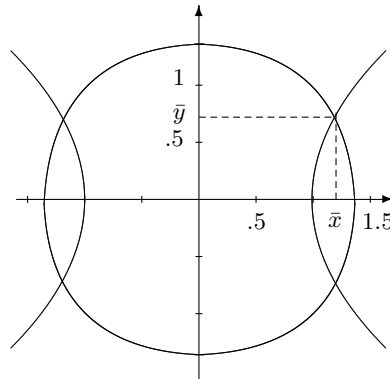



Figura 3.15

Da Figura 3.15, observamos que o sistema dado admite 4 soluções, uma em cada quadrante. Vamos aqui determinar apenas a que se encontra no 1º quadrante. O ponto de interseção das duas equações é a solução $(\bar{x}, \bar{y})$ procurada. Analisando a Figura 3.15, vemos que $(\bar{x}, \bar{y})$ está nas vizinhanças do ponto $(1.2, 0.7)$. Tomemos então: $(x_0, y_0) = (1.2, 0.7)$.

Calculamos primeiramente as derivadas parciais:

$$\begin{aligned} f_x &= 2\,x\ , & f_y &= 2\,y\ , \\ g_x &= 2\,x\ , & g_y &= -2\,y\ . \end{aligned}$$

Assim:

$$\begin{aligned} f(x_0, y_0) &= f(1.2, 0.7) = (1.2)^2 + (0.7)^2 - 2 = -0.07\ , \\ g(x_0, y_0) &= g(1.2, 0.7) = (1.2)^2 - (0.7)^2 - 1 = -0.05\ , \\ f_x(x_0, y_0) &= f_x(1.2, 0.7) = 2 \times (1.2) = 2.4 = g_x(x_0, y_0)\ , \\ g_x(x_0, y_0) &= g_x(1.2, 0.7) = 2 \times (0.7) = 1.4 = -g_y(x_0, y_0)\ . \end{aligned}$$

Então, usando (3.14), obtemos:

$$\begin{aligned} x_1 &= 1.2 - \left[\frac{f g_y - g f_y}{f_x g_y - f_y g_x}\right]_{(1.2, 0.7)} \\ &= 1.2 - \left[\frac{(-0.07)(-1.4) - (-0.05)(1.4)}{-(2.4)(1.4) - (2.4)(1.4)}\right] = 1.225\ , \\ y_1 &= 0.7 - \left[\frac{g f_x - f g_x}{f_x g_y - f_y g_x}\right]_{(1.2, 0.7)} \\ &= 0.7 - \left[\frac{(-0.05)(2.4) - (-0.07)(2.4)}{-(2.4)(1.4) - (2.4)(1.4)}\right] = 0.7071. \end{aligned}$$

Calculando o erro relativo:

$$\left|\frac{x_1 - x_0}{x_1}\right| \simeq 0.02 \ \text{ e } \left|\frac{y_1 - y_0}{y_1}\right| \simeq 0.01$$

observamos que ambos são maiores que 10^{-3}. Assim, devemos fazer nova iteração. Calculemos então:

$$\begin{aligned} f(x_1, y_1) &= f(1.225, 0.7071) = (1.225)^2 + (0.7071)^2 - 2 = -0.000615 \\ g(x_1, y_1) &= g(1.225, 0.7071) = (1.225)^2 - (0.7071)^2 - 1 = -0.000635\ . \\ f_x(x_1, y_1) &= f_x(1.225, 0.7071) = 2 \times (1.225) = 2.45 = g_x(x_1, y_1)\ , \\ g_x(x_1, y_1) &= g_x(1.225, 0.7071) = 2 \times (0.7071) = 1.4142 = -g_y(x_1, y_1) \end{aligned}$$

Novamente, usando (3.14), segue que:

$$\begin{aligned} x_2 &= 1.225 - \left[\frac{f g_y - g f_y}{f_x g_y - f_y g_x}\right]_{(1.225, 0.7071)} \\ &= 1.225 - \left[\frac{(-0.000615)(-1.4142) - (-0.000635)(1.4142)}{-(2.45)(1.4142) - (2.45)(1.4142)}\right] = 1.2253 \\ y_2 &= 0.7071 - \left[\frac{g f_x - f g_x}{f_x g_y - f_y g_x}\right]_{(1.225, 0.7071)} \\ &= 0.7071 - \left[\frac{(-0.000635)(2.45) - (-0.000615)(2.45)}{-(2.45)(1.4142) - (2.45)(1.4142)}\right] = 0.7070. \end{aligned}$$

Calculando o erro relativo:

$$\left|\frac{x_2 - x_1}{x_2}\right| \simeq 0.0002 \text{ e } \left|\frac{y_2 - y_1}{y_2}\right| \simeq 0.0001$$

vemos que estes são menores que 10^{-3}. Assim, a solução do sistema dado é $(\bar{x}, \bar{y}) = (1.2252, 0.7070)$ com $\epsilon < 10^{-3}$.

Exercícios:

3.20 - *Usando o método de Newton determine, com precisão de 10^{-3}, uma raiz para cada um dos seguintes sistemas não lineares:*

$$i) \left\{ \begin{array}{l} 3x^2y - y^3 = 4 \\ x^2 + xy^3 = 9 \end{array} \right. \text{ com } (x_0, y_0) = (2; 2.5).$$

$$ii) \left\{ \begin{array}{l} x^2 + y^2 - 1 = 0 \\ x^2 + y^2 + \frac{1}{2} = 0 \end{array} \right. \text{ com } (x_0, y_0) = (1; 3).$$

$$iii) \left\{ \begin{array}{l} (x-1)^2 + y^2 = 4 \\ x^2 + (y-1)^2 = 4 \end{array} \right. \text{ com } (x_0, y_0) = (2; 1).$$

3.7 Equações Polinomiais

Embora as equações polinomiais possam ser resolvidas por qualquer dos métodos iterativos discutidos previamente, elas surgem tão frequentemente que recebem um tratamento especial. Em particular, apresentaremos alguns algoritmos eficientes para a determinação de raízes isoladas de polinômios, sejam elas reais ou complexas.

Seja

$$\begin{aligned} P(x) &= a_n\, x^n + a_{n-1}\, x^{n-1} + \ldots + a_1\, x + a_0 \\ &= \textstyle\sum_{i=0}^{n} a_i x^i, \quad a_n \neq 0. \end{aligned} \tag{3.15}$$

um polinômio de grau n com $a_n \neq 0$. Então, os seguintes resultados são válidos para $P(x)$:

a) $P(x)$ possui, pelo menos, uma raiz.

b) $P(x)$ possui, exatamente, n raízes, desde que uma raiz de multiplicidade k seja considerada k vezes.

c) Se os valores numéricos de dois polinômios de grau $\leq n$ coincidem para mais do que n valores distintos de x, os polinômios são idênticos.

d) Se $x_1, x_2, \ldots, x_n$ forem as raízes de $P(x)$, então $P(x)$ pode ser expresso univocamente na forma fatorada:

$$P(x) = a_n(x - x_1)(x - x_2) \ \ldots \ (x - x_n) \ .$$

e) Se os coeficientes a_k, $(k = 0, 1, \ldots, n)$ forem reais, e se $a + bi$ for uma raiz complexa de $P(x)$, então $a - bi$ será também uma raiz de $P(x)$.

3.7.1 Determinação de Raízes Reais

Inicialmente deduziremos um algoritmo para a determinação das raízes reais de polinômios. Consideraremos apenas polinômios contendo coeficientes reais. Em qualquer método iterativo para determinação de uma raiz de um polinômio, teremos que calcular, frequentemente, o valor numérico do polinômio para um determinado número real. Portanto, é importante realizar esse cálculo de uma forma tão precisa quanto possível. Por exemplo, usando o método de Newton, temos:

$$x_{k+1} = x_k - \frac{P(x_k)}{P'(x_k)} .$$

Para medir a eficiência dos algoritmos para calcular o valor do polinômio num ponto, usemos a seguinte notação:

- μ = tempo de processamento de uma multiplicação,
- α = tempo de processamento de uma adição,

Se $P(x)$ é calculado pela fórmula (3.15), então devemos calcular as potências de x fazendo $x^k = x \times x^{k-1}$, os quais requerem $(n-1)\mu$; termos da forma $a_k x^{n-k}$, os quais requerem $n\mu$; e a soma dos termos, os quais requerem $n\alpha$. Assim, nessa maneira de cálculo, o total é $(2n-1)\mu + n\alpha$. Além disso, quase a mesma quantidade é requerida se $P'(x)$ é calculado por esse método.

Em vista da simplicidade do problema, é surpreendente que exista um algoritmo que calcula $P(x), P'(x)$ e também as derivadas de ordem superior de $P(x)$, caso se deseje, com uma quantidade muito inferior de tempo de processamento. Esse algoritmo, chamado de **Algoritmo de Briot- Ruffini-Horner**, é obtido escrevendo a fórmula para $P(x)$ da seguinte maneira: (Vamos considerar $n = 4$, para simplicidade.)

$$\begin{aligned} P(x) &= a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \\ &= (((a_4x + a_3)x + a_2)x + a_1)x + a_0 . \end{aligned}$$

Desse modo temos que o tempo de processamento requerido é: $4\mu + 4\alpha$. Assim, de um modo geral, para um polinômio de grau n podemos formular o algoritmo da seguinte maneira: Dados $a_n, a_{n-1}, \ldots, a_0$, calcular $b_n, b_{n-1}, \ldots, b_0$, de acordo com:

$$b_n = a_n , \quad b_{n-k} = x b_{n-k+1} + a_{n-k} , \quad k = 1, 2, \ldots, n . \tag{3.16}$$

Portanto, $b_0 = P(x)$ = valor de P em x. Assim, $\bar{x}$ é uma raiz de $P(x)$ se e somente se, no algoritmo de Briot-Riffini-Horner, formado com o número $\bar{x}$, resultar que $b_0 = 0$. Observe que o tempo de processamento requerido agora é: $n\mu + n\alpha$.

Vamos aplicar agora a b_k o mesmo algoritmo que aplicamos a a_k. Fazendo isso, obtemos números c_k de acordo com:

$$c_n = b_n , \quad c_{n-k} = x c_{n-k+1} + b_{n-k} , \quad k = 1, 2, \ldots, n-1 . \tag{3.17}$$

Para nossa surpresa, $c_1 = P'(x)$, e assim o valor da derivada do polinômio em x é obtida, com tempo de processamento igual a $(n-1)(\mu+\alpha)$. A prova analítica de que $c_1 = P'(x)$, é feita por diferenciação da relação de recorrência dada por (3.16), lembrando que b_k é função de x enquanto que os a_k não. Assim, derivando (3.16), obtemos:

$$b'_n = 0 , \quad b'_{n-k} = x b'_{n-k+1} + b_{n-k+1} , \quad k = 1, 2, \ldots, n .$$

Vemos que $b'_{n-1} = b_n$, e que as quantidades $c_k = b'_{k-1}$ são idênticas aos c_k definidos por (3.17). Portanto desde que $b_0 = P(x)$, segue que : $c_1 = b'_0 = P'(x)$.

Seja $x = z$, então os valores de $P(z)$, fórmulas (3.16), e $P'(z)$, fórmulas (3.17), podem ser obtidos através do esquema prático:

Seja $P(x) \; = \; a_n \; x^n \; + \; a_{n-1} \; x^{n-1} \; + \ldots + \; a_1 \; x \; + \; a_0$. Então:

	a_n	a_{n-1}	a_{n-2}	$\ldots$	a_2	a_1	a_0
	$\downarrow$	$+$	$+$		$+$	$+$	$+$
z		zb_n	zb_{n-1}	$\ldots$	zb_3	zb_2	zb_1
	b_n	b_{n-1}	b_{n-2}	$\ldots$	b_2	b_1	b_0
	$\downarrow$	$+$	$+$		$+$	$+$	
z		zc_n	zc_{n-1}	$\ldots$	zc_3	zc_2	
	c_n	c_{n-1}	c_{n-2}	$\ldots$	c_2	c_1	

com $\mathbf{b_0} = P(z)$ e $\mathbf{c_1} = P'(z)$.

Note que o esquema prático acima quando utilizado para calcular apenas o valor do polinômio num ponto é o conhecido algoritmo de Briot-Ruffini. O esquema de Briot-Ruffini-Horner, na verdade, fornece o valor de $\dfrac{P'(z)}{1!}$, e pode ser continuado para obtenção de $\dfrac{P''(z)}{2!}$, $\dfrac{P'''(z)}{3!}$, etc. (ver [Henrice, 1977]).

Assim, quando $f(x)$ é um polinômio, o método de Newton, fórmula (3.9), pode ser expresso como:

$$x_{k+1} \; = \; x_k \; - \; \frac{b_0(x_k)}{c_1(x_k)} \; , \tag{3.18}$$

onde $b_0(x_k)$ e $c_1(x_k)$ representam, respectivamente, o valor do polinômio e da derivada do polinômio avaliados em x_k.

Vamos assumir agora que z é um zero de $P(x)$. Se $P(z) = 0$ então $b_0 = 0$. Afirmamos então que:
Os números $b_n, b_{n-1}, \ldots, b_1$ são os coeficientes do polinômio $Q(x)$, obtido da divisão de $P(x)$ pelo fator linear $x - z$, isto é:

$$Q(x) \; = \; b_n \; x^{n-1} \; + \; b_{n-1} \; x^{n-2} \; + \; \ldots \; + \; b_1 = \frac{P(x)}{x - z} \; .$$

De fato,

$$\begin{aligned} (b_n \; x^{n-1} \; & + \; b_{n-1} \; x^{n-2} \; + \; \ldots \; + \; b_1)(x - z) \\ & = \; b_n \; x^n + (b_{n-1} - z \; b_n) \; x^{n-1} \; + \; \ldots \\ & + \; (b_1 - z \; b_2) \; x \; + \; (b_0 - z \; b_1) \\ & = \; a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0 = P(x) \; . \end{aligned}$$

onde usamos a fórmula de recorrência dada por (3.16), com x substituido por z.

Assim, se z é uma raiz de $P(x)$, podemos escrever que:

$$P(x) \; = \; (x - z) Q(x) \; ,$$

e portanto concluímos que qualquer raiz de $Q(x)$ é, também, uma raiz de $P(x)$. Isto nos permite operar com um polinômio de grau $n - 1$, ou seja , com $Q(x)$, para calcular as raízes subsequentes de $P(x)$. Esse processo recebe o nome de **Deflação**. Usando esse processo evitamos que um mesmo zero seja calculado várias vezes.

Exemplo 3.14 - *Determinar todas as raízes de:*

$$P(x) \; = \; x^3 + 2 \; x^2 - 0.85 \; x - 1.7 \; ,$$

com precisão de 10^{-2}, usando o método de Newton, para o cálculo da primeira raiz positiva.

Solução: Seja $y_1 = x^3$ e $y_2 = -\ 2\ x^2\ +\ 0.85\ x\ +\ 1.7$. Plotando ambas as curvas no mesmo gráfico, obtemos:

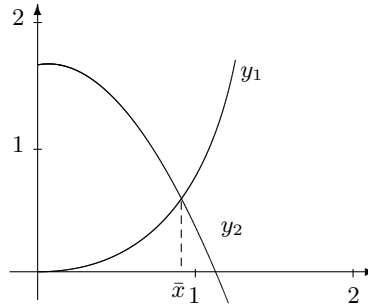



Figura 3.16

Vemos então que $\bar{x}$, está nas vizinhanças de 0.9. Assim, seja $x_0\ =\ 0.9$. Calculemos inicialmente $P(0.9)$ e $P'(0.9)$, usando o algoritmo de Briot-Ruffini-Horner. Portanto:

	1	2	-0.85	-1.7
0.9		0.9	2.61	1.584
	1	2.9	1.76	-0.1164
0.9		0.9	3.42	
	1	3.8	5.18	

Portanto, usando (3.18), segue que:

$$x_1\ =\ 0.9\ -\ \frac{b_0(0.9)}{c_1(0.9)}\ \Rightarrow\ x_1\ =\ 0.9\ -\ \frac{-0.1164}{5.18}\ \Rightarrow\ x_1\ =\ 0.9224\ .$$

Calculando o erro relativo,

$$\left|\frac{x_1 - x_0}{x_1}\right|\ \simeq\ 0.02\ ,$$

vemos que o mesmo é maior que 10^{-2}. Assim devemos fazer nova iteração.

	1	2	-0.85	-1.7
0.9224		0.9224	2.6956	1.7024
	1	2.9224	1.8456	0.0024
0.9224		0.9224	3.5464	
	1	3.8448	5.392	

Logo:

$$x_2\ =\ 0.9224\ -\ \frac{b_0(0.9224)}{c_1(0.9224)} \Rightarrow\ x_2\ =\ 0.9224\ -\ \frac{0.0024}{5.392} \Rightarrow\ x_2\ =\ 0.9220\ .$$

Calculando o erro relativo:

$$\left|\frac{x_2 - x_1}{x_2}\right|\ \simeq\ 0.0004\ ,$$

vemos que este é menor que 10^{-2}, e assim $\bar{x}\ =\ 0.9220$ é uma raiz de $P(x)$ com a precisão exigida. As duas raízes restantes podem ser obtidas, agora, a partir do polinômio do segundo grau: $Q(x) = b_3x^2 + b_2x + b_1$. Aplicando novamente o algoritmo de Briot-Ruffini-Horner, obtemos:

	1	2	−0.85	−1.7
0.9220		0.9220	2.6941	1.7002
	1	2.9220	1.8441	0.0002

e assim podemos escrever que:

$$Q(x) \; = \; x^2 \; + \; 2.9220 \; x \; + \; 1.8441 \; .$$

Usando a fórmula que nos fornece as raízes de uma equação do segundo grau, obtemos que as outras duas raízes de $P(x)$ são: $\bar{x} = -0.9235$ e $\bar{x} = -1.9985$.

Exercícios

3.21 - *Calcular* $P(5)$ *e* $P'(5)$ *para o polinômio:* $P(x) = x^5 - 3x^4 + 2x^2 - 3x + 5$.

3.22 - *Determinar todas as raízes do polinômio:* $P(x) = x^3 - 5x^2 - x + 5 = 0$*, com precisão de* 10^{-2}*, usando o método de Newton e o algoritmo de Briot-Ruffini-Horner para o cálculo da primeira raiz.*

3.23 - *Use o método das secantes e o algoritmo de Briot-Ruffini para determinar a única raiz negativa da equação* $f(x) = x^3 - 2x^2 - x + 2 = 0$*, com precisão de* 10^{-2}.

3.24 - *A equação* $f(x) = x^3 - 0.5 = 0$ *possui uma raiz entre* 0.5 *e* 1.0*. usando o método Regula Falsi e o algoritmo de Briot-Ruffini determinar essa raiz com precisão de* 10^{-2}.

3.7.2 Determinação de Raízes Complexas

O método de Newton pode ser usado também para calcular as raízes complexas de polinômios. Neste caso entretanto devemos usar aritmética complexa. Veremos aqui como determinar as raízes complexas de um polinômio usando aritmética real.

Se $P(x)$ é um polinômio da forma (3.15) com coeficientes reais, as raízes complexas ocorrem, então, em pares conjugados, e, correspondendo a cada par de raízes complexas conjugadas, há um fator quadrático de $P(x)$ da forma:

$$x^2 \; - \; \alpha \, x \; - \; \beta \; ,$$

onde α e β são números reais. Consideremos, primeiramente, a divisão de um polinômio $P(x)$ de grau $n > 2$ por um fator quadrático. É claro que, em geral, podemos expressar $P(x)$ na forma:

$$P(x) \; = \; \left(x^2 \; - \alpha \, x \; - \; \beta\right) Q(x) \; + \; b_1(x-\alpha) \; + \; b_0 \; , \tag{3.19}$$

onde $Q(x)$ é um polinômio de grau $n-2$, que representamos na forma:

$$Q(x) \; = \; b_n \; x^{n-2} \; + \; b_{n-1} \; x^{n-3} \; + \; \ldots \; + b_2 \; , \tag{3.20}$$

e $b_1(x-\alpha) + b_0$ é o resto.

É conhecido da teoria dos polinômios que $x^2 - \alpha \; x - \beta$ será um divisor exato de $P(x)$ se e somente se, $b_1 = b_0 = 0$. Quando $b_1 = b_0 = 0$, a expressão (3.19) torna-se:

$$P(x) \; = \; \left(x^2 \; - \; \alpha \; x \; - \; \beta\right) \; Q(x) \; .$$

Portanto as raízes de $x^2 - \alpha \; x - \beta$ e as raízes de $Q(x)$, serão, também, raízes de $P(x)$. Nosso objetivo é então obter coeficientes α e β, de tal forma que $x^2 - \alpha x - \beta$ seja um divisor exato de $P(x)$, pois teremos duas raízes a partir do fator quadrático e as demais poderemos obter através do polinômio $Q(x)$. Para

determinarmos os coeficientes $b_k,\ k = 0, 1, \ldots, n$ em (3.19) para valores arbitrários de α e β, expandimos o lado direito da igualdade (3.19). Assim:

$$\begin{array}{rcl} P(x) & = & x^2 \left(b_n\ x^{n-2}\ +\ b_{n-1}\ x^{n-3}\ + \ldots +\ b_2\right) \\ & - & \alpha x \left(b_n\ x^{n-2}\ +\ b_{n-1}\ x^{n-3}\ + \ldots +\ b_2\right) \\ & - & \beta\ \ \left(b_n\ x^{n-2}\ +\ b_{n-1}\ x^{n-3}\ + \ldots +\ b_2\right) \\ & + & b_1(x-\alpha)\ +\ b_0 \\ & = & b_n\ x^n\ +\ (b_{n-1} - \alpha\ b_n)\, x^{n-1}\ +\ (b_{n-2} - \alpha\ b_{n-1} - \beta\ b_n)\, x^{n-2} \\ & + & \ldots\ + (b_1 - \alpha\ b_2 - \beta\ b_3)\, x\ +\ (b_0 - \alpha\ b_1 - \beta\ b_2)\,. \end{array}$$

Igualando esses coeficientes aos de $P(x)$ em (3.15) e reagrupando os termos, obtemos as fórmulas de recorrência:

$$\begin{array}{rcl} b_n & = & a_n\ , \\ b_{n-1} & = & a_{n-1}\ +\ \alpha\ b_n\ , \\ b_{n-2} & = & a_{n-2}\ +\ \alpha\ b_{n-1}\ +\ \beta\ b_n\ , \\ \vdots & & \\ b_1 & = & a_1\ +\ \alpha\ b_2\ +\ \beta\ b_3\ , \\ b_0 & = & a_0\ +\ \alpha\ b_1\ +\ \beta\ b_2\ . \end{array} \tag{3.21}$$

Os números $b_n, b_{n-1}, \ldots, b_2$ são os coeficientes do polinômio $Q(x)$.

Esquema Prático para o cálculo de $b_k,\ k = 0, 1, \ldots, n$

Seja $P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0$. Então:

	a_n	a_{n-1}	a_{n-2}	$\ldots$	a_2	a_1	a_0
		$+$	$+$		$+$	$+$	$+$
α	$\downarrow$	αb_n	αb_{n-1}	$\ldots$	αb_3	αb_2	αb_1
			$+$		$+$	$+$	$+$
β	$\downarrow$	$\downarrow$	βb_n	$\ldots$	βb_4	βb_3	βb_2
	b_n	b_{n-1}	b_{n-2}	$\ldots$	b_2	b_1	b_0

Em (3.21) b_1 e b_0 são, logicamente, funções de α e β. Em geral, para uma escolha arbitrária de α e β, eles não se anularão. Encontrar o fator quadrático que seja divisor exato de $P(x)$ equivale a resolver o sistema de equações não lineares:

$$\left\{ \begin{array}{rcl} b_1\ (\alpha, \beta) & = & 0 \\ b_0\ (\alpha, \beta) & = & 0 \end{array} \right. \tag{3.22}$$

Se (α_0, β_0) forem aproximações das raízes $(\bar{\alpha}, \bar{\beta})$ de (3.22), podemos tentar resolver esse sistema pelo método de Newton para funções de duas variáveis. A correção $(\delta\alpha_0)$ de (α_0, β_0) onde:

$$\delta\alpha_0\ =\ \alpha_1\ -\ \alpha_0 \quad \text{e} \quad \delta\beta_0 = \beta_1 - \beta_0\ ,$$

pode ser encontrada solucionando-se o sistema:

$$\left\{ \begin{array}{rcl} \dfrac{\partial b_1}{\partial \alpha}\ \delta\alpha_0\ +\ \dfrac{\partial b_1}{\partial \beta}\ \delta\beta_0 & = & -b_1\,(\alpha_0, \beta_0) \\ \\ \dfrac{\partial b_0}{\partial \alpha}\ \delta\alpha_0\ +\ \dfrac{\partial b_0}{\partial \beta}\ \delta\beta_0 & = & -b_0\,(\alpha_0, \beta_0) \end{array} \right. \tag{3.23}$$

onde as derivadas parciais devem ser calculadas em (α_0, β_0). Uma vez que não podemos expressar b_1 e b_0, explicitamente, como funções de α e β, não podemos calcular explicitamente as derivadas. Bairstow propôs um método simples para calcular numericamente estas derivadas parciais.

Para obter $\dfrac{\partial b_1}{\partial \alpha}$ e $\dfrac{\partial b_0}{\partial \alpha}$ derivamos (3.21) em relação a α, tendo em mente que os a_k são constantes e que os b_k são todos funções de α, exceto b_n. Portanto:

$$
\begin{aligned}
\frac{\partial b_n}{\partial \alpha} &= 0 \;, \\
\frac{\partial b_{n-1}}{\partial \alpha} &= b_n \;, \\
\frac{\partial b_{n-2}}{\partial \alpha} &= b_{n-1} + \alpha \frac{\partial b_{n-1}}{\partial \alpha} \;, \\
\frac{\partial b_{n-3}}{\partial \alpha} &= b_{n-2} + \alpha \frac{\partial b_{n-2}}{\partial \alpha} + \beta \frac{\partial b_{n-1}}{\partial \alpha} \;, \\
\ldots\ldots \\
\frac{\partial b_{n-1}}{\partial \alpha} &= b_2 + \alpha \frac{\partial b_2}{\partial \alpha} + \frac{\partial b_3}{\partial \alpha} \;, \\
\frac{\partial b_0}{\partial \alpha} &= b_1 + \alpha \frac{\partial b_1}{\partial \alpha} + \beta \frac{\partial b_2}{\partial \alpha} \;.
\end{aligned}
\tag{3.24}
$$

Fazendo $c_{k+1} = \dfrac{\partial b_k}{\partial \alpha}$, $k = n-1, n-2, \ldots, 1, 0$, temos que (3.24) pode ser expresso da seguinte maneira:

$$
\begin{aligned}
c_n &= b_n \;, \\
c_{n-1} &= b_{n-1} + \alpha c_n \;, \\
c_{n-2} &= b_{n-2} + \alpha c_{n-1} + \beta c_n \;, \\
c_{n-3} &= b_{n-3} + \alpha c_{n-2} + \beta c_{n-1} \;, \\
&\vdots \\
c_2 &= b_2 + \alpha c_3 + \beta c_4 \;, \\
c_1 &= b_1 + \alpha c_2 + \beta c_3 \;.
\end{aligned}
\tag{3.25}
$$

Comparando (3.25) com (3.21) vemos que os c_k são obtidos a partir dos b_k, da mesma forma como os b_k foram obtidos a partir dos a_k (exceto que não existe o termo c_0). Além disso, as derivadas requeridas são:

$$
\frac{\partial b_0}{\partial \alpha} = c_1 \;, \quad \frac{\partial b_1}{\partial \alpha} = c_2 \;. \tag{3.26}
$$

Para obter $\dfrac{\partial b_1}{\partial \beta}$, $\dfrac{\partial b_0}{\partial \beta}$, derivamos (3.21) em relação a β, tendo em mente que os a_k são constantes e

que os b_k são todos funções de β, exceto b_n e b_{n-1}. Portanto:

$$
\begin{aligned}
\frac{\partial b_n}{\partial \beta} &= \frac{\partial b_{n-1}}{\partial \beta} = 0 \ , \\
\frac{\partial b_{n-2}}{\partial \beta} &= bn \ , \\
\frac{\partial b_{n-3}}{\partial \beta} &= b_{n-1} + \alpha \frac{\partial b_{n-2}}{\partial \beta} \ , \\
\frac{\partial b_{n-4}}{\partial \beta} &= b_{n-2} + \alpha \frac{\partial b_{n-3}}{\partial \beta} + \frac{\partial b_{n-2}}{\partial \beta} \ , \\
&\ldots\ldots \\
\frac{\partial b_1}{\partial \beta} &= b_3 + \alpha \frac{\partial b_2}{\partial \beta} + \beta \frac{\partial b_3}{\partial \beta} \ , \\
\frac{\partial b_0}{\partial \beta} &= b_2 + \alpha \frac{\partial b_1}{\partial \beta} + \beta \frac{\partial b_2}{\partial \beta} \ .
\end{aligned} \tag{3.27}
$$

Fazendo $d_{i+2} = \dfrac{\partial b_i}{\partial \beta}, \ i = n-2, n-3, \ldots, 1, 0$, temos que (3.27) pode ser escrito como:

$$
\begin{aligned}
d_n &= b_n \ , \\
d_{n-1} &= b_{n-1} + \alpha d_n \ , \\
d_{n-2} &= b_{n-2} + \alpha d_{n-1} + \beta d_n \ , \\
c_{n-3} &= b_{n-3} + \alpha d_{n-2} + \beta d_{n-1} \ , \\
&\vdots \\
d_3 &= b_3 + \alpha d_4 + \beta d_5 \ , \\
d_3 &= b_3 + \alpha d_4 + \beta d_5 \ .
\end{aligned} \tag{3.28}
$$

Comparando (3.28) com (3.25) vemos que $d_k = c_k, \ k = 2, 3, \ldots, n$. Portanto:

$$
\frac{\partial b_0}{\partial \beta} = d_2 = c_2 \ , \quad \frac{\partial b_1}{\partial \beta} = d_3 = c_3 \ . \tag{3.29}
$$

Assim, usando (3.26) e (3.30), as equações (3.23) empregadas para a determinação das correções $\delta\alpha_0, \delta\beta_0$, tornam-se:

$$
\left\{
\begin{array}{l}
c_2 \ \delta\alpha_0 + c_3 \ \delta\beta_0 = - \ b_1 (\alpha_0, \beta_0) \\
\\
c_1 \ \delta\alpha_0 + c_2 \ \delta\beta_0 = - \ b_0 (\alpha_0, \beta_0)
\end{array}
\right. \tag{3.30}
$$

Esse método para a determinação de um fator quadrático de um polinômio e as correspondentes raízes é chamado **Método de Newton-Bairstow**.

O método de Newton-Bairstow se constitui num poderoso e eficiente algoritmo para o cálculo das raízes complexas de polinômios. Poderoso porque converge quadraticamente e eficiente porque fornece um algoritmo simples para a obtenção das derivadas parciais requeridas. Sua maior deficiência é que muitas vezes é difícil selecionar adequadamente as aproximações iniciais (α_0, β_0) a fim de garantir a convergência. Entretanto, podemos obter (α_0, β_0), usando o algoritmo Q-D, (ver próxima seção). Observe que o método de Newton-Bairstow pode ser utilizado para obter as raízes reais (desde que uma raiz real pode ser considerada uma raiz complexa cuja parte imaginária é zero) de polinômios com a vantagem de se conseguir duas raízes de cada vez.

Exemplo 3.15 - *Calcular todas as raízes da equação polinomial:*

$$P(x) \; = \; x^4 - 2x^3 + 4x^2 - 4x + 4 = 0$$

pelo método de Newton-Bairstow, iniciando com $(\alpha_0, \beta_0) = (1, -1)$.

Solução: Primeiramente calculamos os b_k e os c_k. Assim:

	1	-2	4	-4	4
1		1	-1	2	-1
-1			-1	1	-2
	1	-1	2	-1	1
1		1	0	1	
-1			-1	0	
	1	0	1	0	

Resolvemos, então o sistema:

$$\begin{cases} 1.\ \delta\alpha_0 \; + \; 0.\ \delta\beta_0 \; = \; 1 \\ 0.\ \delta\alpha_0 \; + \; 1.\delta\beta_0 \; = \; -1 \end{cases}$$

cuja solução é: $\delta\alpha_0 = 1$ e $\delta\beta_0 = -1$.

Assim:

$$\begin{array}{l} \alpha_1 \; = \; \alpha_0 \; + \; \delta\alpha_0 \; \Rightarrow \; \alpha_1 \; = \; 2 \; , \\ \beta_1 \; = \; \beta_0 \; + \; \delta\beta_0 \; \Rightarrow \; \beta_1 \; = \; -2 \; . \end{array}$$

Repetimos, então, o processo com $(\alpha_1, \beta_1) \; = \; (2, -2)$. Logo:

	1	-2	4	-4	4
2		2	0	4	0
-2			-2	0	-4
	1	0	2	0	0

Obtemos então $b_1 = b_0 = 0$. Logo, $x^2 - \alpha\, x - \beta \; = \; x^2 - 2\,x + 2$ é um divisor exato de $P(x)$. Portanto as raízes de $x^2 - 2x + 2$ são também raízes de $P(x)$. Mas,

$$P(x) = \left(x^2 - 2x + 2\right) \; Q(x) \; ,$$

onde:

$$Q(x) = x^2 + 2 \; .$$

Assim, as raízes de $Q(x)$ são também raízes de $P(x)$.

Logo as raízes de $P(x)$ são: $\begin{cases} 1 \; \pm \; i \\ \\ \pm \; \sqrt{2}\; i. \end{cases}$

Exercícios

3.25 - *Dividir o polinômio* $x^6 - 3x^5 + 4x^2 - 5$ *por* $x^2 - 3x + 1$, *e excrevê-lo na forma (3.19).*

3.26 - *Usar o método de Newton-Bairstow para determinar as raízes de* $P(x) = x^3 - 6x^2 + 9x - 4$, *partindo da divisão de* $P(x)$ *por* $x^2 - x$.

3.7.3 Algoritmo Quociente-Diferença

Os métodos de Newton e Newton-Bairstow para determinação de zeros de polinômios são eficientes se conhecemos, respectivamente, uma aproximação inicial suficientemente próxima da raiz, ou uma aproximacão inicial adequada para o fator quadrático.

Nessa seção apresentaremos um método numérico que determina os zeros de um polinômio sem conhecer aproximações iniciais, mesmo que as raízes sejam complexas. Tal método, conhecido como **Algoritmo Quociente-Diferença**, ou simplesmente **Algoritmo Q-D**, é um esquema devido a Rutishauser, que fornece simultaneamente aproximações para todos os zeros de um polinômio, sejam eles reais ou complexos. Maiores detalhes sobre o algoritmo Q-D, podem ser encontrados em [Henrici, 1964] ou em [Albrecht, 1973].

Seja $P(x)$ um polinômio da forma (3.15), isto é:

$$P(x) \;=\; a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0 \;.$$

Vamos considerar que $P(x)$ é um polinômio de grau $n \geq 1$, com $a_k \neq 0, k = 0, 1, \ldots, n$.

A partir de $P(x)$ construímos linhas de termos q e e, começando a tabela calculando a primeira linha de $q's$ e a segunda linha de $e's$, da seguinte maneira:

$$q_0^{(1)} \;=\; -\,\frac{a_{n-1}}{a_n}\;; \quad q_0^{(k)} \;=\; 0\;, \quad k = 2, \ldots, n\;,$$

$$e_0^{(k)} \;=\; \frac{a_{n-(k+1)}}{a_{n-k}}\;, \quad k = 1, 2, \ldots, n-1\;; \quad e_0^{(1)} = e_0^{(n)} = 0\;.$$

Assim as duas primeiras linhas da tabela são:

$e^{(0)}$	$q^{(1)}$	$e^{(1)}$	$q^{(2)}$	$e^{(2)}$	$q^{(3)}$	$\ldots$	$e^{(n-1)}$	$q^{(n)}$	$e^{(n)}$
	$-\frac{a_{n-1}}{a_n}$		0		0	$\ldots$		0	
0		$\frac{a_{n-2}}{a_{n-1}}$		$\frac{a_{n-3}}{a_{n-2}}$		$\ldots$		$\frac{a_0}{a_1}$	0

As novas linhas de $q's$ serão calculadas através da equação:

$$\text{novo } q^{(k)} \;=\; e^{(k)} - e^{(k-1)} + q^{(k)}\;, \quad k = 1, 2, \ldots, n\;, \tag{3.31}$$

usando os termos das linhas e e q acima. Note que nessa equação o novo q é igual ao *e à direita menos e à esquerda mais q acima*. As novas linhas de $e's$ são calculadas pela equação:

$$\text{novo } e^{(k)} \;=\; \frac{q^{(k+1)}}{q^{(k)}}e^{(k)}\;, \quad k = 1, 2, \ldots, n\;, \quad e^{(0)} = e^{(n)} = 0\;, \tag{3.32}$$

onde o novo e é igual ao *q à direita sobre q à esquerda vezes e acima.*

Utilizamos sucessivamente as fórmulas (3.31) e (3.32) até que os $e's$ tendam a zero. Quando isso ocorrer, os valores de q aproximam os valores das raízes, se estas forem reais. Se o polinômio tiver um par de raízes complexas conjugadas, um dos $e's$ não tenderá a zero mas *flutuará* em torno de um valor. Nesse caso devemos montar um fator quadrático da forma: $x^2 - rx - s$, do seguinte modo: a soma dos dois valores de q, um de cada lado do valor de e em questão, aproximará o valor de r e o produto do valor de q acima e à esquerda vezes o valor de q abaixo e à direita aproximará o valor de $-s$. Fazendo $x^2 - rx - s = 0$, determinamos as raízes complexas. Caso semelhante vale para raízes de multiplicidade 2. Daremos a seguir exemplo.

Exemplo 3.16 - *Usando o algoritmo Q-D obter todas as raízes do polinômio:*

$$P(x) \;=\; x^4 - 6x^3 + 12x^2 - 19x + 12 \; .$$

Solução: Aplicando o algoritmo Q-D, obtemos a tabela a seguir, onde após calcularmos as duas primeiras linhas indicamos com setas como calcular os novos $q's$ e $e's$:

$e^{(0)}$	$q^{(1)}$	$e^{(1)}$	$q^{(2)}$	$e^{(2)}$	$q^{(3)}$	$e^{(3)}$	$q^{(4)}$	$e^{(4)}$
	6.000		0		0		0	
0		−2.000		−1.583		−0.632		0
	4.000		0.417		0.951		0.632	
0		−0.208		−3.610		−0.420		0
	3.792		−2.985		4.141		1.052	
0		0.164		5.008		−0.107		0
	3.956		1.859		−0.974		1.159	
0		0.077		−2.624		0.127		0
	4.033		−0.842		1.777		1.032	
0		−0.016		5.538		0.074		0
	4.017		4.712		−3.687		0.958	
0		−0.019		−4.333		−0.019		0
	3.998		0.398		0.627		0.977	
0		−0.002		−6.826		−0.030		0
	4.000		**−6.426**		7.423		1.007	
0		0.003		7.885		−0.004		0
	4.003		**1.456**		**−0.466**		**1.010**	

Observe que na tabela acima $q^{(1)}$ está convergindo para 4 e $q^{(4)}$ está convergindo para 1. Assim 4.004 e 1.010 são aproximações para duas das raízes de $P(x)$. Agora, desde que $e^{(2)}$ não está convergindo para zero, $q^{(2)}$ e $q^{(3)}$, representam o fator quadrático: $x^2 - rx - s$, onde:

$$\begin{aligned} r &= 1.456 + (-0.466) \;=\; 0.990 \; , \\ s &= (-6.426) \times (-0.466) \;=\; -2.995 \; . \end{aligned}$$

Portanto igualando o fator quadrático a zero, isto é, fazendo:

$$x^2 - rx - s = x^2 - 0.990x + 2.995 = 0,$$

obtemos que: $0.495 \pm 1.6568i$ são aproximações para as outras duas raízes de $P(x)$. Podemos então escrever que:

$$P(x) \simeq (x - 4.004)(x - 1.010)(x - (0.495 + 1.6568i))(x - (0.495 - 1.6568i)) \; .$$

É claro, como já dissemos, que os valores encontrados são aproximações para as raízes de $P(x)$. Se desejarmos o resultado com mais casas decimais corretas, podemos aplicar o **Algoritmo Q-D versão Newton**: aplica-se o algoritmo Q-D para obter aproximações para as raízes, e usando o método de Newton ou Newton-Bairstow refina-se a solução até obtê-la com a precisão desejada. Esta versão de Newton faz com que o algoritmo seja praticamente livre de erros de arredondamento.

Observações: O algoritmo Q-D não se aplica se:

a) durante o processo ocorrer algum $q = 0$, (divisão por zero).

b) o polinômio dado tiver raízes nulas, (é exigido que todos os coeficientes sejam diferentes de zero).

c) se existir algum coeficiente igual a zero.

Para entender a observação do item **a)** resolva o primeiro exercício proposto a seguir. Em relação ao item **b)** o processo pode ser aplicado desde que elimenemos do polinômio as raízes nulas antes de aplicá-lo. Em relação ao item **c)**, isto é, se existir algum coeficiente igual a zero basta fazer uma mudança de variável: $z = x - a$, onde a é uma constante real arbitrária. Com a mudança de variável obtemos um polinômio que possui todos os coeficientes diferentes de zero. Aplicamos a esse polinômio o algoritmo Q-D. Determinadas as raízes usamos a mudança de variável para obter os zeros do polinômio dado, isto é: $x = z + a$. Assim:

Exemplo 3.17 - *Dado* $P(x) = 81x^4 - 108x^3 + 24x + 20$, *determinar um polinômio que possua todos os coeficientes diferentes de zero.*

Solução: Temos em $P(x)$ que o coeficiente $a_2 = 0$. Fazemos então a mudança de variável: $z = x - a$. Com essa mudança de variável obteremos um polinômio $P^*(z)$. Observe que tal polinômio é facilmente obtido se desenvolvermos $P(x)$ em série de Taylor em torno do ponto a. De fato, para o polinômio dado, obtemos que:

$$\begin{aligned} P(x) &= P(a) + (x-a)P'(a) + \frac{(x-a)^2}{2!}P''(a) \\ &+ \frac{(x-a)^3}{3!}P'''(a) + \frac{(x-a)^4}{4!}P^{(iv)}(a) \ . \end{aligned}$$

Fazendo $x - a = z$, obtemos:

$$P^*(z) = P(a) + zP'(a) + z^2\frac{P''(a)}{2!} + z^3\frac{P'''(a)}{3!} + z^4\frac{P^{(iv)}(a)}{4!} \ .$$

Os coeficientes do polinômio $P^*(z)$ são obtidos aplicando-se o algoritmo de Briot-Ruffini-Horner, (ver seção 3.7.1). Como o valor de a é arbitrário, em geral, consideramos $a = 1 \rightarrow z = x - 1$. Portanto, para o polinômio dado, devemos calcular o valor do polinômio $P(x)$ e de suas derivadas no ponto $a = 1$. Assim:

	81	−108	0	24	20
1		81	−27	−27	−3
	81	−27	−27	−3	**17**
1		81	54	27	
	81	54	27	**24**	
1		81	135		
	81	135	**162**		
1		81			
	81	**216**			
1	**81**				

Logo:

$$P^*(z) = 81z^4 + 216z^3 + 162z^2 + 24z + 17 \ .$$

Usamos a algoritmo Q-D para calcular as raízes de $P^*(z)$ e a seguir fazendo $x = z + 1$, obtemos as raízes de $P(x)$.

Exercícios

3.27 - *Verifique que não é possível determinar as raízes de* $P(x) = x^2 - 2x + 2$, *usando o algoritmo Q-D.*

3.28 - *Usando o algoritmo Q-D determinar todas as raízes de:*

a) $P(x) = 81x^4 - 108x^3 + 24x + 20$,

b) $P(x) = 128x^4 - 256x^3 + 160x^2 - 32x + 1$.

com duas casas decimais corretas.

3.8 Exercícios Complementares

3.29 - *Mostre que as seguintes equações possuem exatamente uma raiz e que em cada caso a raiz está no intervalo* $[0.5, 1]$.

a) $x^2 + \ln\ x = 0$,

b) $x\ e^x - 1 = 0$.

Determine essas raízes, com duas casas decimais corretas, usando o método da bissecção.

3.30 -*Aplique o método da bisseção e Regula Falsi para calcular a raiz positiva de* $x^2 - 7 = 0$ *com* $\epsilon < 10^{-2}$, *partindo do intervalo inicial* $[2.0, 3.0]$.

3.31 -*Aplique o método da bissecção para resolver:*

a) $e^x - x - 3\ x = 0$,

b) $x^3 + cos\ x = 0$,

obtendo em cada caso a e b (iniciais) graficamente.

3.32 - *O problema: resolva* $f(x) = x + ln\ x = 0$ *pode ser transformado num problema equivalente da forma* $x = \psi(x)$. *Para o processo iterativo definido por* $x_{k+1} = \psi(x)$; *analisar a convergência quando:*

a) $\psi(x) = -ln\ x$,

b) $\psi(x) = e^{-x}$,

no intervalo $[0.5, 0.6]$.

3.33 - *A equação* $x^2 + 5x - 1 = 0$ *tem uma raiz em* $(0, 0.5)$. *Verifique quais dos processos abaixo podem ser usados, com sucesso, para obtê-la:*

a) $x_{k+1} = \dfrac{1 - x_k^2}{5}$,

b) $x_{k+1} = \dfrac{1 - 5\ x_k}{x_k}$,

c) $x_{k+1} = \sqrt{1 - 5\ x_k}$.

3.34 - *A equação $f(x) = e^x - 3\ x^2 = 0$ tem três raízes. Um método iterativo pode ser definido usando a preparação óbvia da equação:*

$$x = \pm\sqrt{\frac{e^x}{3}}\ .$$

i) *Verificar que começando com $x_0 = 0$ haverá convergência para a raiz próxima de -0.5, se o valor negativo for usado e que haverá convergência para a raiz próxima de 1.0, se o valor positivo for usado.*

ii) *Mostrar que a forma acima não converge para a terceira raiz próxima de 4.0, qualquer que seja a aproximação inicial próxima da raiz.*

3.35 - *A fórmula $x_{n+1} = 2\ x_n - a\ x_n^2$ é candidata para se determinar o inverso de um número a; $\frac{1}{a}$. Mostre que se a fórmula converge, então converge para $\frac{1}{a}$ e determine os limites da estimativa inicial x_0 para convergir. Teste suas conclusões nos casos:*

a) $a = 9$ *e* $x_0 = 0.1$.

b) $a = 9$ *e* $x_0 = 1.0$.

3.36 - *Mostre que $x^3 - 2x - 17 = 0$ tem apenas uma raiz real e determine seu valor correto até 2 casas decimais usando o método de Newton.*

3.37 - *A equação $x^3 - 2x - 1 = 0$ possui apenas uma raiz positiva.*

a) *De acordo com o princípio da bissecção, esta raiz positiva deve estar em qual dos intervalos: $(0,1), (1,2), (2,3)$? Por que?*

b) *Se desejássemos também pesquisar as raízes negativas usando intervalos de amplitude $\frac{1}{2}$, até o ponto -2, em que intervalos seriam encontradas tais raízes?*

c) *Obtenha a menor raiz negativa (em módulo), usando o método das Secantes. Trabalhe com arredondamento para 3 casas decimais.*

3.38 - *Usando o método de Newton determine t (real), com erro relativo inferior a 10^{-2}, tal que a matriz:*

$$A\ =\ \begin{pmatrix} 0.5 & 0.2 & t \\ 0.4 & t & 0.5 \\ t & 0.5 & 0.2 \end{pmatrix}\ ,$$

seja singular.

3.39 - *Usando o método de Newton determine, sem efetuar a divisão, o valor numérico de $x = \frac{1}{3}$ com 3 casas decimais corretas, iniciando com $x_0 = 0.3$.*

3.40 - *A solução da equação diferencial:*

$$a_n\ \frac{d^n u(t)}{dt^n}\ +\ a_{n-1}\ \frac{d^{n-1}u(t)}{dt^{n-1}}\ +\ \ldots\ +\ a_1\ \frac{du(t)}{dt}\ =\ 0\ ,$$

é:

$$u(t)\ =\ \sum_{k=1}^{m} q_k(t) e^{\lambda_k t}\ ,$$

onde os λ_k são as raízes distintas do polinômio:

$$P(\lambda) \; = \; a_n \; \lambda^n \; + \; a_{n-1} \; \lambda^{n-1} \; + \; \ldots \; + \; a_1 \; \lambda \; + \; a_0 \; ,$$

chamado polinômio característico da equação diferencial e os q_k são polinômios de grau uma unidade inferior à multiplicidade de λ_k, mas a não ser por isso, arbitrários.

Deseja-se determinar a solução geral da equação diferencial:

$$\frac{d^3u(t)}{dt^3} \; - \; 6 \; \frac{d^2u(t)}{dt^2} \; + \; 6 \; \frac{du(t)}{dt} \; + \; 7 \; u(t) \; = \; 0 \; .$$

Determine a única raiz negativa do polinômio característico pelo método de Newton e o algoritmo de Briot-Ruffini-Horner, com erro inferior a 10^{-3}, e as demais raízes através da equação do 2º grau.

3.41 - *Seja $\bar{x}$ uma raiz da equação $f(x) = 0$. Supomos que $f(x)$, $f'(x)$ e $f''(x)$ sejam contínuas e limitadas num intervalo fechado I contendo $x = \bar{x}$ e que $f'(\bar{x}) = 0$ e $f''(\bar{x}) \neq 0$. (Observe que nestas condições $\bar{x}$ é um zero de multiplicidade 2 de $f(x) = 0$).*

a) *Mostre que o método iterativo definido por:*

$$x_{k+1} \; = \; x_k \; - \; 2 \; \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots$$

converge para a raiz $\bar{x}$ se $x_k \in I$.

b) *O método definido em* **a)** *estende-se para uma raiz de multiplicidade m da seguinte maneira:*

$$x_{k+1} \; = \; x_k \; - \; m \; \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, \ldots$$

Calcular a raiz $\bar{x}$ próxima de 1, da equação:

$$f(x) \; = \; x^4 - 3.1 \; x^3 + 2.52 \; x^2 + 0.432 \; x - 0.864 \; = \; 0 \; ,$$

com erro relativo inferior a 10^{-3}, usando o método descrito acima e sabendo que $f(\bar{x}) = f'(\bar{x}) = f''(\bar{x}) = 0$ e $f'''(\bar{x}) \neq 0$.

3.42 -*Dado o sistema não linear:*

$$\begin{cases} x^3 - 3 \; x \; y^2 + 1 = 0 \\ 3 \; x^2 \; y - y^3 = 0 \end{cases}$$

Determine uma solução com 2 dígitos significativos corretos, iniciando com $(x_0, y_0) = (0.51, 0.85)$, e usando:

a) *método iterativo linear*

b) *método de Newton.*

3.43 - *Mostre que o sistema não linear:*

$$\begin{cases} 3x^2 + y^2 + 9 \; x - y - 12 = 0 \\ x^2 + 36 \; y^2 - 36 = 0 \end{cases}$$

possui exatamente 4 raízes. Determine essas raízes usando o método de Newton, com 2 dígitos significativos corretos, iniciando com $(1, 1), (1, -1), (-4, 1)$ e $(-4, -1)$.

3.44 - *Sejam $C = (1, 0)$ e $D = (0, 1)$. Usando o método de Newton para sistemas não lineares determine o valor de um ponto $P = (x, y)$, com precisão de 10^{-3}, que diste 2 unidades de C e de D, obtendo os valores iniciais necessários através de gráfico.*

3.45 - *Considere o seguinte problema: "dado um polinômio de grau n com coeficientes reais, $P(z)$, onde z é uma variável complexa, determinar uma raiz complexa de $P(z)$, se existir, ou seja resolver a equação $P(z) = 0$". Como $z = x + i\ y$, o polinômio $P(z)$ pode ser escrito na forma:*

$$P(z) \ = \ u(x, y) + i\ v(x, y)$$

Então resolver a equação $P(z) = 0$ é equivalente a resolver o seguinte sistemas de equações:

$$\begin{cases} u(x, y) = 0 \\ v(x, y) = 0 \end{cases}$$

Dada uma aproximação inicial (x_0, y_0) conveniente, podemos resolver este sistema pela extensão do método de Newton (para sistemas não lineares).

Aplique o processo descrito acima para determinar uma aproximação da raiz complexa de $P(z) = z^2 - 2\ z + 3$, tomando como valor inicial $(x_0, y_0) = (1, 1)$.

3.46 - *Dado que: $x^2 - 3.9\ x + 4.8$ é um fator aproximado de $x^4 - 4\ x^3 + 4\ x^2 + 4\ x - 5 = 0$, use o método de Newton-Bairstow para melhorar a aproximação.*

3.47 - *Considere a matriz:*

$$A \ = \ \begin{pmatrix} 1 & -2 & 0 & 0 \\ t^2 & 8 & 0 & 0 \\ 0 & 0 & t & 2 \\ 0 & 0 & 2 & t \end{pmatrix} .$$

Sabendo que o determinante de A é o polinômio $P(t) = 2\ t^4 - 32$, determine todos os valores de t, com erro inferior a 10^{-3}, que tornem a matriz singular. Utilize o método de Newton-Bairstow e use como fator quadrático inicial $x^2 - 0.0001\ x - 3.999$. Trabalhe com arredondamento para 4 casas decimais.

3.48 - *Usando o algoritmo Q-D versão Newton (para a raiz real) e Newton-Bairstow (para as raízes complexas), determine todos os zeros de $P(x) = x^3 - x^2 + 2x - 2$, com 4 casas decimais corretas.*

3.9 Problemas Aplicados e Projetos

3.1 - *A equação de Kepler, usada para determinar órbitas de satélites é dada por:*

$$M \ = \ x \ - \ E \ sen \ x \ .$$

Dado que $E = 0.2$ e $M = 0.5$, obtenha a raiz da equação de Kepler.

3.2 - *Em problemas de fluxo em tubulações, é frequente precisar resolver a equação: $c_5\ D^5 + c_1\ D + c_0 = 0$. Se $c_5 = 1000$, $c_1 = -3$ e $c_0 = 9.04$, determine uma primeira raiz usando o método de Newton e então aplique o método de Newton-Bairstow para determinar as demais raízes.*

3.3 - *Um amplificador eletrônico com acoplamento R - C com três estágios em cascata tem uma resposta a um degrau unitário de tensão dada pela expressão:*

$$g(T) = 1 - (1 + T + \frac{T^2}{2})e^{-T} \ ,$$

onde $T = \dfrac{t}{RC}$ *é uma unidade de tempo normalizada. O tempo de subida de um amplificador é definido como o tempo necessário para sua resposta ir de* 10% *a* 90% *de seu valor final. No caso, como* $g(\infty) = 1$ *é necessário calcular os valores de* T *para os quais*

$$g = 0.1 \quad e \quad g = 0.9$$

ou seja resolver as equações:

$$\begin{aligned} 0.1 &= 1 - (1 + T + \frac{T^2}{2})e^{-T} \ . \\ 0.9 &= 1 - (1 + T + \frac{T^2}{2})e^{-T} \ . \end{aligned}$$

Chamando de $T_{0.1}$ *o valor obtido de* T *na* 1^a *equação e* $T_{0.9}$ *o valor obtido de* T *na* 2^a *equação, calcular o tempo de subida.*

3.4 - *A Figura 3.17 represente o fluxo de água em um canal aberto.*

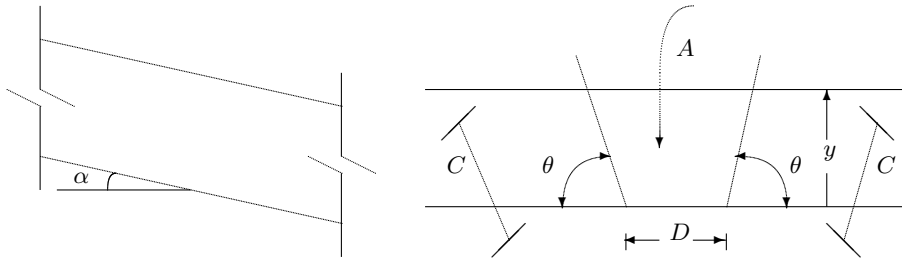



Figura 3.17

Uma relação empírica para o fluxo é a equação de Chez-Manning:

$$Q \ = \ \frac{1.49}{E} \ A \ R^{\frac{2}{3}} \ S^{\frac{1}{2}} \ ,$$

onde:

- Q*- fluxo em* $m^3/seg.$,
- E *- coeficiente de atrito determinado experimentalmente, valendo entre* 0.025 *e* 0.035 *para a maioria dos canais e rios* ,
- A *- área da secção transversal do canal* ,
- R *- raio hidráulico que é definido como a razão entre a área* A *e o perímetro* $2\ C + D$,
- α *- inclinaçào do canal* ($S \ = \ sen\ \alpha$).

a) *Para um canal retangular* ($\theta \ = \ 90^o$), *sendo conhecidos* Q, E, S, D, *verificar que* y *é a solução da equação:*

$$\left[\left(\frac{1.49}{E} \right)^3 \ D^5 \ S^{\frac{3}{2}} \right] \ y^5 \ - \ 4\ Q^3\ y^2 \ - \ 4\ Q^3\ D\ y \ - \ Q^3\ D^2 \ = \ 0 \ ,$$

a qual tem apenas uma raiz positiva.

b) *Encontre as profundidades y do canal correspondente a duas estações A e B, cujos dados estão tabelados a seguir:*

Estação	D	S	E	Q
(A)	20.0	0.0001	0.030	133.0
(B)	21.5	0.0001	0.030	122.3

Em cada caso determinar incialmente intervalo contendo a raiz.

3.5 - *A Figura 3.18 corresponde a um cabo uniforme, como por exemplo uma linha de transmissão suspensa em dois apoios e sob a ação de seu próprio peso.*

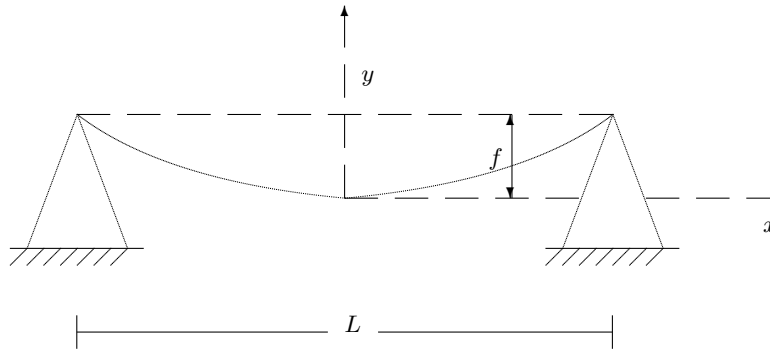



Figura 3.18

A curva correspondente é uma catenária, cuja equação é dada por:

$$y = \frac{T_0}{\mu} \left(cosh \frac{\mu\, x}{T_0} - 1 \right) ,$$

onde:

- T_0 - *tração no cabo em* $x = 0$,
- μ - *peso por unidade de comprimento do cabo* .

Em $x = \frac{L}{2},\ y = f,$ *logo:*

$$f = \frac{T_0}{\mu} \left(cosh \frac{\mu\, L}{2\, T_0} - 1 \right) .$$

O comprimento S do cabo é dado por:

$$f = \frac{2\, T_0}{\mu} \left(senh \frac{\mu\, L}{2\, T_0} \right) .$$

Resolva então o seguinte problema: Um cabo de telefone pesando 1.5 Kgf/m *está simplesmente apoiado em dois pontos cuja distância é de* 30 *metros. Para um comprimento de cabo de* 33 *metros qual é o valor da flecha f?*

3.6 - *A equação:*

$$tg\left(\frac{\theta}{2}\right) = \frac{sen\ \alpha\ cos\ \alpha}{\frac{g\ R}{v^2} - cos^2\ \alpha} ,$$

permite calcular o ângulo de inclinação, α, em que o lançamento do míssil deve ser feito para atingir um determinado alvo. Na equação acima,

- α *- ângulo de inclinação com a superfície da Terra com a qual é feita o lançamento do míssil ,*
- g *- aceleração da gravidade* $\simeq 9.81\ m/s^2$ *,*
- R *- raio da Terra* $\simeq 6371000\ m$ *,*
- v *- velocidade de lançamento do míssil,* m/s *,*
- θ *- ângulo (medido do centro da Terra) entre o ponto de lançamento e o ponto de impacto desejado ,*

Resolva o problema considerando: $\theta = 80^o$ *e* v *tal que* $\frac{v^2}{gR} = 1.25$*, ou seja, aproximadamente* $8.840\ m/s$*.*

3.7 - *Quando um capacitor carregado é ligado com uma resistência R, um processo de descarga do capacitor ocorre. Durante este processo, uma variável no tempo é estabelecida no circuito. Sua variação com o tempo se dá de forma decrescente e exponencial, de acordo com a expressão:*

$$F(t) = I = \frac{Q_0}{RC}\ e^{-\frac{T}{RC}} ,$$

onde I *é a corrente,* Q_0 *é a carga inicial do capacitor,* C *sua capacitância,* R *a resistência e* T *o parâmetro tempo.*

Definindo $G(t) = F(t) - I$*, o instante* T *em que* $G(t) = 0$*, corresponde àquele em que a corrente* I *percorre o circuito.*

Determinar T *nos seguintes casos:*

a) $I = 0.83$ *Ampére,* $Q_0 = 7$ *coulomb,* $R = 3$ *Ohms,* $C = 2$ *Farad;*

b) $I = 0.198$ *Ampére,* $Q_0 = 20$ *coulomb,* $R = 9$ *Ohms,* $C = 11$ *Farad.*

3.8 - *Uma loja de eletrodomésticos oferece dois planos de financiamento para um produto cujo preço a vista é R$ 162,00:*

- *Plano A: entrada de R$ 22,00 + 9 prestações iguais de R$ 26,50,*
- *Plano B: entrada de R$ 22,00 + 12 prestações de R$ 21,50.*

Qual dos dois planos apresenta a menor taxa de juros, sendo portanto melhor para o consumidor?

Observação: *Sabe-se que a equação que relaciona os juros* **(J)** *e o prazo* **(P)** *com o valor financiado (***VF** *= preço à vista - entrada) e a prestação mensal* **PM** *é dada por:*

$$\frac{1 - (1+J)^{-P}}{J} = \frac{VF}{PM} . \quad (3.33)$$

a) *Fazendo* $x = 1 + J$ *e* $k = \frac{VF}{PM}$*, verificar que a equação (3.33) se transforma em:*

$$f(x) = kx^{P+1} - (k+1)x^P + 1 = 0 . \quad (3.34)$$

b) *Escrever a equação (3.34) para o problema proposto e encontrar um intervalo contendo a raiz positiva $\neq$ 1.*

3.9 - *Um dos elfos de Valfenda, o grande arqueiro Glorfindel, disparou uma flecha em direção a cidade de Bri para cair na cabeça de Cevado Carrapicho, dono da estalagem do Pônei Saltitante. O rei Elessar, de Gondor, viu o fato em sua pedra vidente. Junto porém aparecia o seguinte escrito:*

$$37.104740 + 3.15122t - \frac{2t^2}{2} = 0 \; .$$

Elessar desesperado, pois adorava a cerveja da estalagem, queria salvar Cevado Carrapicho (fabricante da cerveja) a qualquer custo; mas apesar de toda sua sabedoria, não entendia o que significavam aqueles números. Como ele podia ver o futuro em sua pedra, correu até uma gruta e escreveu numa parede o seguinte:

"Por favor, quem souber o que significa:

$$37.104740 + 3.15122t - \frac{2t^2}{2} = 0 \; ,$$

me ajude!"

Elessar esperou por um minuto e colocou sua pedra de forma a ver os escritos e verificou que logo abaixo da sua escrita aparecia:

$$\text{“}t = -4.71623 \quad ou \quad t = 7.86745 \; ,$$

que deve ser o tempo de alguma coisa, em horas ou minutos."

Elessar levou algum tempo para traduzir a escrita, mas logo correu para ajudar Cevado, pois se ele estivesse no alvo depois de 7 horas e 52 minutos seria acertado. Elessar conseguiu chegar a tempo e salvou Cevado da morte certa, e comemorou com sua tão amada cerveja...

Dezenas de milhares de anos depois....

Eric estava vasculhando uma gruta quando encontrou escritos junto a rabiscos. Ele percebeu que os rabiscos eram runas élficas, e que aquilo era um pedido de ajuda.

Graças a Deus e aos Anjos, Eric estava com seu notebook na mochila, e tinha um programa chamado Raízes que seu irmão havia instalado para resolver alguns problemas. Depois de alguns segundos tentando entender como eram entrados os dados, ele obteve:

$$\text{“}t = -4.71623 \quad ou \quad t = 7.86745 \; ,''$$

e pensou, isso deve ser alguma coisa, em horas ou minutos....

a) *Resolva o problema proposto, e obtenha pelo método de Newton a raiz positiva.*

b) *Obter a raíz negativa, usando o polinômio do primeiro grau obtido no esquema de Briot-Riffini-Horner.*

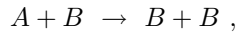
3.10 - *Na engenharia química, reatores do tipo PFR são frequentemente usados para converter reagentes em produtos. Sabe-se que a eficiência de conversão às vezes pode ser melhorada reciclando uma fração do produto como mostrado na Figura 3.19 a seguir:*

Figura 3.19

A taxa de reciclo é definida por:

$$R = \frac{\text{volume do fluido que retorna ao reator}}{\text{volume do fluido que sai do reator}}$$

Supondo que estamos processando um reagente A a fim de gerar um reagente B, segundo a expressão autocatalítica:

$$A + B \rightarrow B + B \; ,$$

pode-se mostrar que a taxa ótima de reciclo satisfaz a equação:

$$\ln\left[\frac{1 + R(1 - x_A)}{R(1 - x_A)}\right] = \frac{R + 1}{R[1 + R(1 - x_a)]} \; , \qquad (3.35)$$

onde x_A é a fração de reagente A que é convertido no produto B. A taxa ótima de reciclo corresponde ao reator de menor tamanho possível necessário para se atingir o nível de conversão desejado. Determine as razões de reciclo necessárias para se minimizar o tamanho do reator, resolvendo a equação (3.35) para as seguintes frações de conversão (x_A), do reagente A no produto B:

i) $x_A = 0.99$,

ii) $x_A = 0.995$,

iii) $x_A = 0.999$,

iv) $x_A = 0.9999$,

v) $x_A = 0.99999$.

3.11 - *Suponha que tenhamos um circuito temporizador 555 como mostra a Figura 3.20:*

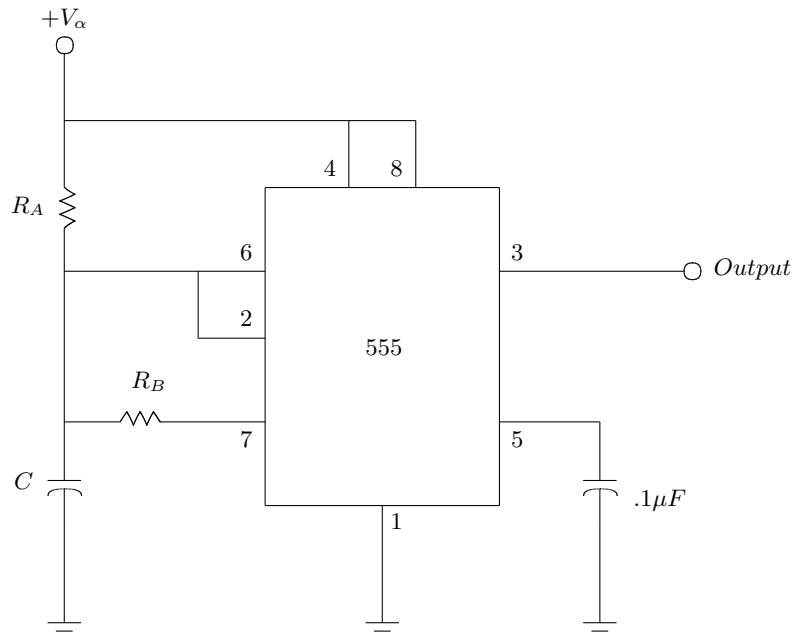



Figura 3.20

cuja onda de saída é da forma:

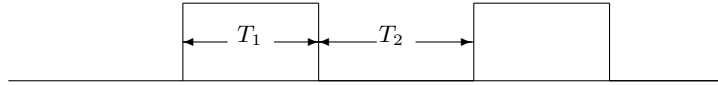



Figura 3.21

com

$$T_1 + T_2 = \frac{1}{f},$$

onde f é a frequência, e o ciclo de trabalho CT, é dado por:

$$CT = \frac{T_1}{T_1 + T_2} \times 100\%.$$

Pode-se mostrar que:

$$T_1 = R_A C ln(2) \ ,$$

$$T_2 = -\frac{R_A \ R_B \ C}{R_A + R_B} \times ln\left(\left|\frac{R_A - 2 \ R_B}{2 \ R_A - R_B}\right|\right) \ .$$

Dado que $R_A = 8.670$, $C = 0.1 \times 10^{-6}$, $T_2 = 1.4 \times 10^{-4}$, determine R_B, T_1, f, e o ciclo de trabalho CT.

3.12 - *Um tanque de vaporização flash é alimentado com $F moles/h$ por uma corrente de gás natural de n componentes, como mostrado na Figura 3.22:*

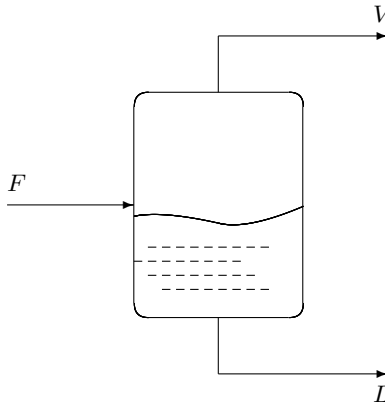



Figura 3.22

As correntes de líquido e vapor são designadas por L e $V moles/h$, respectivamente. As frações molares dos componentes na alimentação, nas correntes de vapor e de líquido são designadas por z_i,y_i e x_i, respectivamente. Assumindo equilíbrio líquido-vapor em estado estacionário, segue que:

$$F \ = \ L + V \ , \tag{3.36}$$

$$z_i F \;=\; x_i L + y_i V \;, \tag{3.37}$$

$$K_i \;=\; \frac{y_i}{x_i} \;, \quad i = 1, 2, \ldots, n \;. \tag{3.38}$$

onde: (3.36) é o balanço global, (3.37) é o balanço individual , (3.38) é a relação de equilíbrio, e, K_i *á a constante de equilíbrio para o i-ésima componente na pressão e temperatura do tanque. Das equações acima e do fato de* $\sum_{i=1}^{n} \; x_i = \sum_{i=1}^{n} \; y_i = 1$*, mostra-se que:*

$$\sum_{i=1}^{n} \; \frac{z_i \; (k_i - 1)}{V(K_i - 1) + F} \;=\; 0 \;. \tag{3.39}$$

Supondo que $F = 1000$ *moles/h, calcule o valor de* V*, com duas casas decimais corretas, resolvendo a equação (3.40), para a corrente de gás natural, à temperatura de* 120^oF *e pressão de 1600 psia, para cada um dos componentes da tabela a seguir:*

Componentes	i	z_i	K_i
Dióxido de Carbono	*1*	*0.0046*	*1.65*
Metano	*2*	*0.8345*	*3.09*
Etano	*3*	*0.0381*	*80.72*
Propano	*4*	*0.0163*	*0.39*
Isobutano	*5*	*0.0050*	*0.21*
n-Butano	*6*	*0.0074*	*0.175*
Pentanos	*7*	*0.0287*	*0.093*
Hexanos	*8*	*0.0220*	*0.065*
Heptanos	*9*	*0.0434*	*0.036*

Para cada valor de V*, calcule os valores de* L*, de* x_i *e de* y_i*.*

3.13 - *Lee and Duffy (A. I. Ch. E Journal, 1976) relacionaram o fator de atrito para escoamentos de partículas fibrosas em suspensão com o número de Reynolds, pela seguinte equação empírica:*

$$\frac{1}{\sqrt{f}} = \left(\frac{1}{k}\right) ln(RE\sqrt{f}) + \left(14 - \frac{5.6}{k}\right) .$$

Nesta relação f *é o fator de atrito,* RE *é o número de Reynolds e* k *é uma constante determinada pela concentração de partículas em suspensão. Para uma suspensão de* 0.08% *de concentração temos que* $k = 0.28$*. Determine o valor de* f *quando* $RE = 3750$*.*

3.14 - *Muitas equações de estado foram desenvolvidas para descrever as relações entre pressão,* P*, volume molar,* V*, e temperatura,* T*, de gases. Uma das equações mais utilizadas é equação de Beattie-Bridgeman:*

$$P \;=\; \frac{RT}{V} + \frac{\beta}{V^2} + \frac{\gamma}{V^3} + \frac{\delta}{V^4} \;, \tag{3.40}$$

onde R *é a constante universal dos gases e* β, γ *e* δ *são parâmetros característicos do gás em estudo. O segundo, terceiro e quarto termos de (3.40), podem ser vistos como correções da lei dos gases ideais,* $PV = RT$*, para o comportamento não ideal de um gás. Os parâmetros* β, γ *e* δ *são definidos por:*

$$\beta \;=\; RTB_0 - A_0 - \frac{Rc}{T^2} \;,$$

$$\gamma = -RTB_0b + A_0a - \frac{RcB_0}{T^2} ,$$
$$\delta = \frac{RB_0bc}{T^2} ,$$

onde: A_0, B_0, a, b *e* c *são constantes determinadas experimentalmente e são diferentes para cada gás. Dados os valores de pressão,* P*, temperatura ,* T*, e das constantes* R, A_0, B_0, a, b, c *é possível determinar o volume molar de qualquer gás resolvendo a equação (3.40), usando como estimativa inicial para o volume molar a lei dos gases ideais:* $V_0 = \frac{RT}{P}$*. Para o gás metano, tem-se que:* $A_0 = 2.2769, B_0 = 0.05587, a = 0.01855, b = -0.01587$ *e* $c = 12.83 \times 10^4$*. Considere temperaturas de* 0^o *e* 200^o *e as seguintes pressões em atm:* $1, 2, 5, 20, 40, 60, 80, 120, 140, 160, 180$ *e* 200*. Com esses dados, determine:*

a) *o volume molar do gás metano, com precisão de* 10^{-6}*,*

b) *o fator de compressibildade* z*, onde* $z = \frac{PV}{RT}$*,*

b) *compare os seus resultados com valores experimentais do fator de compressibilidade para o metano de* 0^o *e* 200^o*, apresentados na Figura 3.23:*

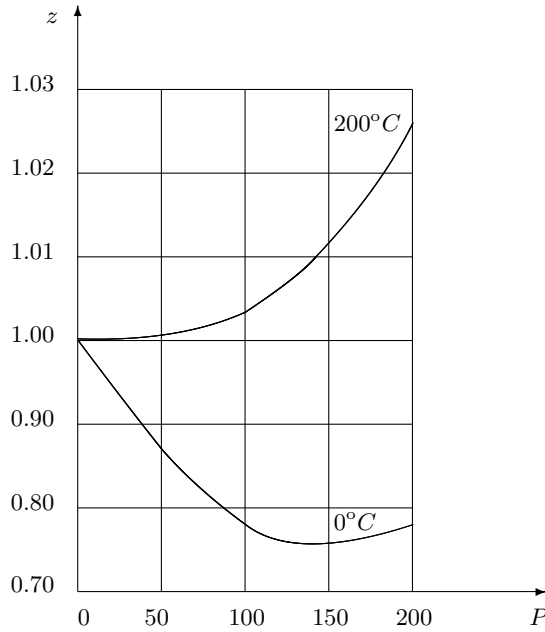



Figura 3.23

3.15 - *Considere o sistema diferencial de segunda ordem:*

$$\begin{cases} x'' + x + 2\ y' + y = f(t) \\ x'' - x + y = g(t) \\ x(0) = x'(0) = y(0) = 0 \end{cases}$$

Para resolvê-lo pelo método da transformada de Laplace, torna-se necessário fatorar a expressão (determinar as raízes):

$$(S^2 + 1)(S) - (2S + 1)(S^2 - 1) ,$$

tal que as frações parciais possam ser usadas no cálculo da transformada inversa. Determine esses fatores (raízes).

3.16 - *Um método muito eficiente para integração numérica de uma função é o chamado método de quadratura de Gauss. No desenvolvimento das fórmulas para este método é necessário calcular os zeros de uma família de polinômios ortogonais. Uma família importante de polinômios ortogonais é a de Legendre. Encontre os zeros do polinômio de Legendre de grau 6(seis):*

$$P_6(x) = \frac{1}{48}(693\ x^6 - 945\ x^4 + 315\ x^2 - 15)\ .$$

Observação: *Todos os zeros dos polinômios de Legendre são menores do que 1(um) em módulo e são simétricos em relação a origem.*

3.17 - *Considere um circuito de polarização que consiste de uma bateria com uma tensão* $V_B = 2.0\ V$ *e um resistor R de* $50\ \Omega$ *em série, conectado a um diodo semicondutor de estado sólido como mostrado na Figura 3.24.*

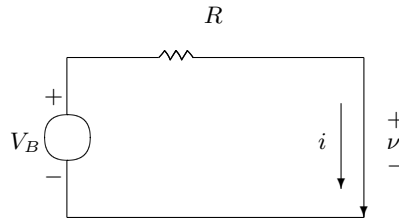



Figura 3.24

As características operacionais na gama normal de operação de um diodo são determinadas pela equação relacionando suas variáveis terminais de tensão e corrente. Se tomarmos ν *e* i *como sendo estas variáveis e escolhermos as direções de referência relativas mostradas, a equação relacionando estas variáveis é dada por:*

$$i\ =\ I_s\ \left(e^{\frac{q\nu}{k\ t}} - 1\right)\ , \tag{3.41}$$

onde:

- I_s *é a intensidade de corrente de saturação reversa. Esta é a corrente máxima que flui quando o diodo é polarizado em reverso, ou seja, quando* $\nu\ <<\ 0$. *Ela é função do material usado na confecção do diodo, do grau de lubrificação e das técnicas de fabricação particulares. Um valor típico para um diodo de silício em temperatura ambiente é* 10^{-9} *Ampére ,*
- k *é a constante de Boltzmann, que tem o valor:* 1.38047×10^{-23} *joule/°K,*
- t *é a temperatura absoluta em °K na qual o diodo é operado,*
- q *é a carga do elétron que tem o valor:* 1.6020310^{-19} *coulomb.*

Em temperaturas ambientes normais, o valor do termo $\frac{q}{kt}$ *é aproximadamente* 40.

Podemos agora proceder à solução do circuito de polarização, ou seja, encontrar os valores de ν *e* i. *Para isso, basta aplicar a lei das tensões de Kirchoff ao circuito, obtendo assim:*

$$V_B\ =\ i\ R + \nu\ . \tag{3.42}$$

Substituindo em (3.42) os valores de V_B *e R, e usando a relação dada por (3.41), obtem-se uma equação não linear em* ν. *Resolvendo-se esta equação, o valor da corrente de polarização i, é facilmente obtido.*

3.18 - *Num escoamento turbulento em uma rede de tubulação interconectada, a razão de escoamento V de um nó para outro é proporcional à raiz quadrada da diferença entre as pressões nos nós. Para a rede da figura a seguir é solicitado determinar a pressão em cada nó.*

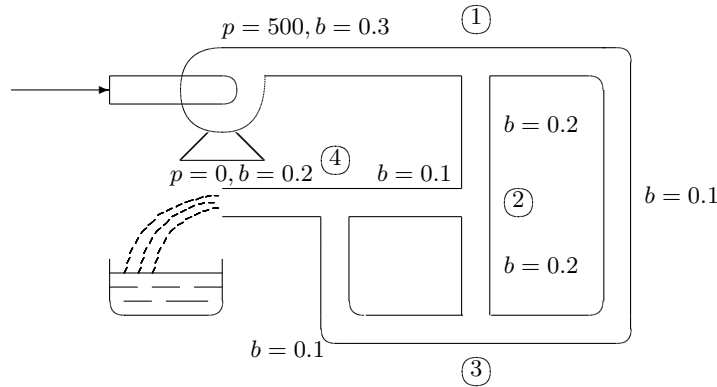



Figura 3.25

Os valores de b representam fatores de condutância na relação:

$$v_{ij} = b_{ij} \sqrt{(p_i - p_j)} \ .$$

As equações para as pressões em cada nó são então dadas por:

$$\begin{aligned}
\text{nó 1} \quad &: 0.3\sqrt{500 - p_1} \ = \ 0.2\sqrt{p_1 - p_2} + 0.2\sqrt{p_1 - p_3} \ , \\
\text{nó 2} \quad &: 0.2\sqrt{p_1 - p_2} \ = \ 0.1\sqrt{p_2 - p_4} + 0.2\sqrt{p_2 - p_3} \ , \\
\text{nó 3} \quad &: 0.1\sqrt{p_1 - p_3} \ = \ 0.2\sqrt{p_3 - p_2} + 0.1\sqrt{p_3 - p_4} \ , \\
\text{nó 4} \quad &: 0.1\sqrt{p_2 - p_4} + 0.1\sqrt{p_3 - p_4} \ = \ 0.2\sqrt{p_4 - 0} \ .
\end{aligned}$$

onde estamos assumindo que $p_1 > p_3$; se isso não for verdadeiro é necessário modificar as equações. Resolva o sistema dado pelo método de Newton.

Capítulo 4

Solução de Sistemas Lineares: Métodos Exatos

4.1 Introdução

Uma variedade de problemas de engenharia pode ser resolvido através da análise linear; entre eles podemos citar: determinação do potencial em redes elétricas, cálculo da tensão na estrutura metálica da construção civil, cálculo da razão de escoamento num sistema hidráulico com derivações, previsão da concentração de reagentes sujeitos à reações químicas simultâneas. O problema matemático em todos estes casos se reduz ao problema de resolver um sistema de equações simultâneas. Também as encontramos, quando estudamos métodos numéricos para resolver problemas de equações diferenciais parciais, pois estes requerem a solução de um conjunto de equações.

A solução de um conjunto de equações é muito mais difícil quando as equações são não lineares. Entretanto a maioria das aplicações envolve somente equações lineares, muito embora quando o sistema é de grande porte devemos escolher o método numérico adequadamente para preservar a máxima precisão.

Antes de desenvolvermos alguns métodos específicos, discutiremos o que queremos dizer com uma solução e as condições sob as quais a solução existe, pois não adianta tentar obter uma solução se não há nenhuma.

Uma equação é *linear* se cada termo contém não mais do que uma variável e cada variável aparece na primeira potência. Por exemplo, $3x + 4y - 10z = -3$ é linear, mas $xy - 3z = -3$ não é, pois o primeiro termo contém duas variavéis. Também $x^3 + y - z = 0$ não é linear, pois o primeiro termo contém uma variável elevada ao cubo.

Vamos considerar n equações lineares com n variáveis (incógnitas) e vamos nos referir a elas como um **Sistema de n Equações Lineares** ou um **Sistema Linear de ordem n**. Uma solução para esse sistema de equações consiste de valores para as n variáveis, tais que quando esses valores são substituídos nas equações, todas elas são satisfeitas simultaneamente.

Por exemplo, o sistema de três equações lineares:

$$\left\{ \begin{array}{rcrcrcr} x & + & y & + & z & = & 1 \\ x & - & y & - & z & = & 1 \\ 2x & + & 3y & - & 4z & = & 9 \end{array} \right.$$

tem a solução $x = 1, y = 1$ e $z = -1$. O leitor pode verificar a validade das três equações substituindo esses valores pelas variáveis no sistema dado.

Observe que o sistema acima pode ser escrito na forma matricial como:

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ 2 & 3 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 9 \end{pmatrix} .$$

De um modo geral um sistema de n equações lineares é escrito como:

$$\left\{ \begin{array}{l} a_{11}\, x_1 \; + \; a_{12}\, x_2 \; + \; a_{13}\, x_3 \ldots \; + \; a_{1n}\, x_n \; = \; b_1 \\ a_{21}\, x_1 \; + \; a_{22}\, x_2 \; + \; a_{23}\, x_3 \ldots \; + \; a_{2n}\, x_n \; = \; b_2 \\ a_{31}\, x_1 \; + \; a_{32}\, x_2 \; + \; a_{33}\, x_3 \ldots \; + \; a_{2n}\, x_n \; = \; b_2 \\ \ldots \; \ldots \\ a_{n1}\, x_1 \; + \; a_{n2}\, x_2 \; + \; an3\, x_3 \ldots \; + \; a_{nn}\, x_n \; = \; b_n \end{array} \right. \tag{4.1}$$

e é representado na forma matricial por:

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} , \tag{4.2}$$

ou simplesmente

$$Ax \; = \; b \; , \tag{4.3}$$

onde A é chamada de matriz dos coeficientes, b é o vetor do termo independente e x é o vetor solução.

Dado um sistema de equações arbitrário, não podemos afirmar sem investigar que há uma solução ou, se houver, que seja única. Como pode ser observado a seguir, há três e apenas três possibilidades de se classificar um sistema linear.

Classificação de um Sistema Linear

A classificação de um sistema linear é feita em função do número de soluções que ele admite, da seguinte maneira:

a) **Sistema Possível ou Consistente**: É todo sistema que possui pelo menos uma solução. Um sistema linear possível é:

(a.1) **determinado** se admite uma única solução, e,

(a.2) **indeterminado** se admite mais de uma solução.

b) **Sistema Impossível ou Inconsistente**: É todo sistema que não admite solução.

O próximo exemplo ilustra a classificação de um sistema linear.

Exemplo 4.1 - *Classificar os seguintes sistemas lineares:*

$$(I) \left\{ \begin{array}{rcrcl} x & + & y & = & 6 \\ x & - & y & = & 2 \end{array} \right.$$

$$(II)\left\{\begin{array}{rcrcl} x & + & y & = & 1 \\ 2\,x & + & 2\,y & = & 2 \end{array}\right.$$

$$(III)\left\{\begin{array}{rcrcl} x & + & y & = & 1 \\ x & + & y & = & 4 \end{array}\right.$$

Solução: Consideremos o sistema (I). A solução é $x = 4$ e $y = 2$; nenhum outro par de valores de x e y satisfaz ambas as equações. Esse sistema é representado geometricamente pela Figura 4.1. Qualquer ponto da reta r_1 tem coordenadas que satisfazem a primeira das equações em (I). Do mesmo modo, todos os pontos em r_2 satisfazem a segunda equação de (I). Os pontos que satisfazem ambas as equações devem localizar-se em ambas as retas. Há somente um ponto assim. As coordenadas desse ponto são a solução que procuramos. Logo, o sistema (I) admite como única solução o par $(4,\ 2)^t$. Portanto (I) é um sistema possível e determinado.

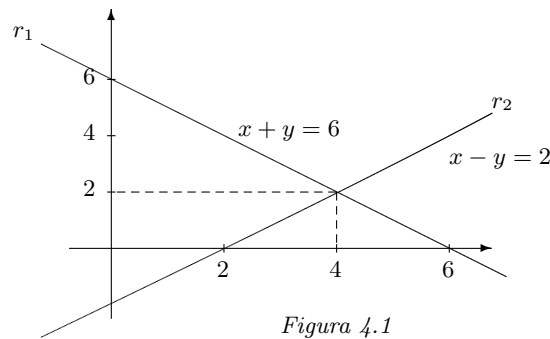



Figura 4.1

Consideremos agora o sistema (II). A Figura 4.2 mostra o gráfico dessas duas retas.

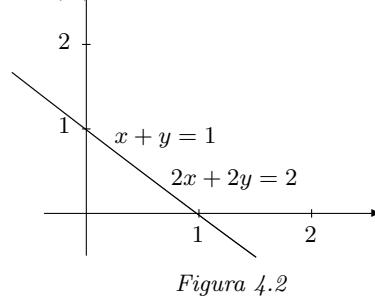
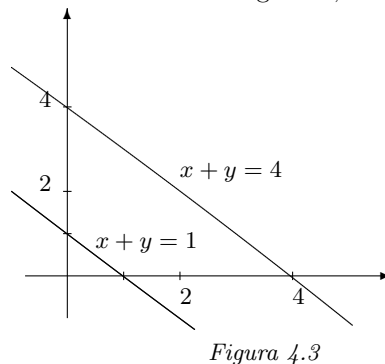
Figura 4.2

Observe que geometricamente as retas $x + y = 1$ e $2\,x + 2\,y = 2$ são coincidentes. Assim, para o sistema (II), temos que os pares $(0,\ 1)^t$; $(1,\ 0)^t$; $(0.5,\ 0.5)^t, \ldots$, são soluções, isto é, o sistema admite infinitas soluções. Logo (II) é um sistema possível e indeterminado. Finalmente, consideremos o sistema (III). Novamente, colocando as duas retas no mesmo gráfico, obtemos a Figura 4.3.

Figura 4.3

Observe que geometricamente as duas retas são paralelas. Assim, para o sistema (III), as duas equações são *contraditórias*, isto é, não é possível que se tenha simultaneamente $x + y = 1$ e $x + y = 4$. Logo (III) é um sistema impossível.

Nosso objetivo aqui será o de desenvolver métodos numéricos para resolver sistemas lineares de ordem n, que tenham solução única. Observe que tais sistemas são aqueles onde a matriz dos coeficientes é não singular, isto é, $det(A) \neq 0$.

Antes de descrevermos em detalhes os métodos de solução, vamos examinar quais os caminhos mais gerais para se chegar a elas.

Métodos numéricos para solução de sistemas de equações lineares são divididos principalmente em dois grupos:

- **Métodos Exatos**: são aqueles que forneceriam a solução exata, não fossem os erros de arredondamento, com um número finito de operações.

- **Métodos Iterativos**: são aqueles que permitem obter a solução de um sistema com uma dada precisão através de um processo infinito convergente.

Assim, os métodos exatos em princípio, ou seja desprezando os erros de arredondamento, produzirão uma solução, se houver, em um número finito de operações aritméticas. Um método iterativo, por outro lado, iria requerer em princípio, um número infinito de operações aritméticas para produzir a solução exata. Assim, um método iterativo tem um erro de truncamento e o exato não tem. Por outro lado, em sistemas de grande porte os erros de arredondamento de um método exato podem tornar a solução sem significado, enquanto que nos métodos iterativos os erros de arredondamento não se acumulam. Veremos entretanto, que ambos são úteis, ambos têm vantagens e limitações.

Neste capítulo estudaremos somente métodos exatos e no Capítulo 5 os métodos iterativos.

Voltemos ao exemplo 4.1. Observando a Figura 4.1 vê-se facilmente que poderíamos traçar infinitos conjuntos de duas retas concorrentes cuja intersecção fosse o par $(4,\ 2)^t$. Cada um desses conjuntos formaria um sistema de duas equações lineares que teriam portanto a mesma solução. Assim definimos:

Definição 4.1 - *Dois sistemas lineares são equivalentes quando admitem a mesma solução.*

Com base na definição 4.1 não fica difícil deduzir que uma maneira de obter a solução de um sistema linear através de métodos numéricos é transformá-lo em outro equivalente cuja solução seja facilmente obtida. Em geral, nos métodos exatos, transformamos o sistema original num sistema equivalente, cuja solução é obtida resolvendo-se sistemas triangulares.

Solução de Sistemas Triangulares

Como já dissemos, resolver sistemas triangulares é muito fácil; entretanto, apresentaremos aqui a solução de tais sistemas com o objetivo de auxiliar a elaboração de projetos que envolvam a resolução dos mesmos.

i) Um sistema linear de ordem n é triangular inferior se tiver a forma:

$$\left\{\begin{array}{lllllll}a_{11}\ x_1 & & & & & = & b_1\\ a_{21}\ x_1 & + & a_{22}\ x_2 & & & = & b_2\\ a_{31}\ x_1 & + & a_{32}\ x_2 & + & a_{33}\ x_3 & = & b_3\\ \ldots\ldots & & \ldots\ldots & & & & \vdots\\ a_{n1}\ x_1 & + & a_{n2}\ x_2 & + & \ldots + a_{nn}\ x_n & = & b_n\end{array}\right.$$

onde $a_{ii} \neq 0,\ i = 1, 2, \ldots, n$. Assim a solução de um sistema triangular inferior é obtida por substituição direta, isto é, determinamos o valor de x_1 na primeira equação; substituímos esse valor na segunda equação e determinamos o valor de x_2 e assim por diante. Algebricamente podemos resolvê-lo pelas fórmulas:

$$\left\{\begin{array}{lll}x_1 & = & \dfrac{b_1}{a_{11}},\\[2ex] x_i & = & \dfrac{b_i - \sum_{j=1}^{i-1} a_{ij}\ x_j}{a_{ii}}\ ,\quad i\ =\ 2, 3, \ldots, n\ .\end{array}\right. \tag{4.4}$$

ii) Um sistema linear de ordem n é triangular superior se tiver forma:

$$\left\{\begin{array}{lllllll}a_{11}\ x_1 & + & a_{12}\ x_2 & + & a_{13}\ x_3\ \ldots + a_{1n}\ x_n & = & b_1\\ & & a_{22}\ x_2 & + & a_{23}\ x_3\ \ldots + a_{2n}\ x_n & = & b_2\\ & & & & a_{33}\ x_3\ \ldots + a_{3n}\ x_n & = & b_n\\ & & & & \ldots\ldots & & \vdots\\ & & & & a_{nn}\ x_n & = & b_n\end{array}\right.$$

onde $a_{ii} \neq 0;\ i = 1, 2, \ldots, n$. Assim a solução de um sistema triangular superior é obtida por retro-substituição, isto é, determinamos o valor de x_n na última equação; substituímos esse valor na penúltima equação e determinamos o valor de x_{n-1} e assim por diante. Algebricamente podemos resolvê-lo pelas fórmulas:

$$\left\{\begin{array}{lll}x_n & = & \dfrac{b_n}{a_{nn}}\ ,\\[2ex] x_i & = & \dfrac{b_i\ -\ \sum_{j=i+1}^{n} a_{ij}\ x_j}{a_{ii}}\ ,\quad i\ =\ n-1, \ldots, 1\ .\end{array}\right. \tag{4.5}$$

Portanto, para resolvermos nosso problema falta explicar como um dado sistema linear de ordem n pode ser transformado num outro equivalente cuja solução seja obtida resolvendo-se sistemas triangulares. Como veremos, os métodos numéricos apresentados nesse capítulo explicam como fazer isso. Antes, porém, daremos algumas definições que julgamos necessárias para um melhor entendimento de tais métodos.

Definição 4.2 - *Uma matriz triangular inferior é uma matriz quadrada $C = (c_{ij})$ tal que $c_{ij} = 0$ para $i < j$. Do mesmo modo, se $c_{ij} = 0$ para $i > j, C$ é uma matriz triangular superior.*

Definição 4.3 *Seja A uma matriz $n \times n$ da forma:*

$$A\ =\ \begin{pmatrix}a_{11} & a_{12} & \ldots & a_{1n}\\ a_{21} & a_{22} & \ldots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n1} & a_{n2} & \ldots & a_{nn}\end{pmatrix}\ . \tag{4.6}$$

Os menores principais de A de ordens $1, 2, \ldots n$ *são definidos pelas sub-matrizes de A:*

$$A_k = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{pmatrix}, \quad k = 1, 2, \ldots n \; .$$

Definição 4.4 - *Uma matriz real simétrica* A, $n \times n$, *é* **positiva definida** *se para todos os menores principais* A_k, *constituído das* k *primeiras linhas e* k *primeiras colunas vale:* $det(A_k) > 0 \;, \quad k = 1, 2, \ldots, n$.

4.2 Decomposição LU

Inicialmente veremos em que condições podemos decompor uma matriz quadrada $A = (a_{ij})$ no produto de uma matriz triangular inferior por uma matriz triangular superior.

Teorema 4.1 - **Teorema LU** - *Seja* $A = (a_{ij})$ *uma matriz quadrada de ordem* n, *e* A_k *o menor principal, constituído das* k *primeiras linhas e* k *primeiras colunas de* A. *Assumimos que* $det(A_k) \neq 0$ *para* $k = 1, 2, \ldots, n-1$. *Então existe uma única matriz triangular inferior* $L = (\ell_{ij})$, *com* $\ell_{11} = \ell_{22} = \ldots = \ell_{nn} = 1$, *e uma única matriz triangular superior* $U = (u_{ij})$ *tal que* $LU = A$. *Além disso,* $det(A) = u_{11}u_{22}\ldots u_{nn}$.

Prova: Para provar esse teorema usaremos indução sobre n.

1) Se $n = 1$, temos que: $a_{11} = 1 \cdot a_{11} = 1 \cdot u_{11}$ unicamente, e assim $A = LU$ onde $L = 1$ e $U = u_{11}$. Além disso, $det(A) = u_{11}$.

2) Assumimos que o teorema é verdadeiro para $n = k-1$, ou seja que toda matriz de ordem $(k-1)$ é decomponível no produto LU nas condições do teorema.

3) Devemos mostrar que a decomposição pode ser feita para uma matriz de ordem $n = k$. Seja então A uma matriz de ordem k. Partimos essa matriz em sub-matrizes da forma:

$$A = \begin{pmatrix} A_{k-1} & r \\ & \\ s^t & a_{kk} \end{pmatrix},$$

onde r e s são vetores coluna, ambos com $k-1$ componentes.

Note que a matriz A_{k-1} é de ordem $k-1$ e satisfaz as hipóteses do teorema. Portanto pela hipótese de indução esta pode ser decomposta na forma $A_{k-1} = L_{k-1}U_{k-1}$. Utilizando as matrizes L_{k-1} e U_{k-1} formamos as seguintes matrizes:

$$L = \begin{pmatrix} L_{k-1} & 0 \\ & \\ m^t & 1 \end{pmatrix}; \quad U = \begin{pmatrix} U_{k-1} & p \\ & \\ 0 & u_{kk} \end{pmatrix},$$

onde m e p são vetores coluna, ambos com $k-1$ componentes. Note que m, p e u_{kk} são desconhecidos. Assim, impondo que a matriz A seja decomponível em LU vamos tentar determiná-los.

Efetuando o produto LU, segue que:

$$LU = \begin{pmatrix} L_{k-1}U_{k-1} & L_{k-1}\, p \\ & \\ m^t\, U_{k-1} & m^t\, p + u_{kk} \end{pmatrix}.$$

Estudemos agora a equação $LU = A$, isto é:

$$\begin{pmatrix} L_{k-1}U_{k-1} & L_{k-1}\ p \\ m^t\ U_{k-1} & m^t\ p + u_{kk} \end{pmatrix} = \begin{pmatrix} A_{k-1} & r \\ s^t & a_{kk} \end{pmatrix} .$$

Da igualdade acima concluímos que:

$$\begin{aligned} L_{k-1}U_{k-1} &= A_{k-1} , \\ L_{k-1}\ p &= r , \\ m^t\ U_{k-1} &= s^t , \\ m^t\ p + u_{kk} &= a_{kk} . \end{aligned}$$

Observe que a primeira equação é válida pela hipótese de indução, e portanto L_{k-1} e U_{k-1} são unicamente determinadas. Além disso, nem L_{k-1} e nem U_{k-1} são singulares (ou A_{k-1} também seria singular, contrariando a hipótese). Assim de:

$$\begin{aligned} L_{k-1}\ p &= r \quad \Rightarrow \quad p = L_{k-1}^{-1}\ r , \\ m^t\ U_{k-1} &= s^t \quad \Rightarrow \quad m^t = s^t\ U_{k-1}^{-1} , \\ m^t\ p + u_{kk} &= a_{kk} \quad \Rightarrow \quad u_{kk} = a_{kk} - m^t\ p . \end{aligned}$$

Portanto p, m e u_{kk} são determinados univocamente nesta ordem, e L e U são determinados unicamente. Finalmente,

$$\begin{aligned} det(A) &= det(L) \cdot det(U) \\ &= 1 \cdot det\,(U_{k-1}) \cdot u_{kk} \\ &= u_{11}u_{22} \ldots\ldots u_{k-1,k-1}u_{kk} , \end{aligned}$$

completando a prova.

Cabe salientar que a decomposição LU fornece um dos algoritmos mais eficientes para o cálculo do determinante de uma matriz.

Esquema Prático para a Decomposição LU

Observe que teoricamente, para obtermos as matrizes L e U, devemos calcular a inversa de L_{k-1} e U_{k-1}. Entretanto na prática podemos calcular L e U simplesmente aplicando a definição de produto e de igualdade de matrizes, isto é, impondo que $LU = A$. Seja então:

$$LU = \begin{pmatrix} 1 & & & & \\ \ell_{21} & 1 & & \bigcirc & \\ \ell_{31} & \ell_{32} & 1 & & \\ \ldots & \ldots & \ldots & \ddots & \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \ldots & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ & u_{22} & u_{23} & \ldots & u_{2n} \\ & & u_{33} & \ldots & u_{3n} \\ & \bigcirc & & \ddots & \vdots \\ & & & & u_{nn} \end{pmatrix} ,$$

e a matriz A como em (4.6).

Para obtermos os elementos da matriz L e da matriz U devemos calcular os elementos das linhas de U e os elementos da colunas de L na seguinte ordem:

1ª linha de U: Fazendo o produto da 1ª linha de L por todas as colunas de U e igualando com os elementos da 1ª linha de A, obtemos ,

$$\begin{array}{rcl} 1 \cdot u_{11} & = & a_{11} \Rightarrow \mathbf{u_{11}} = a_{11} \; , \\ 1 \cdot u_{12} & = & a_{12} \Rightarrow \mathbf{u_{12}} = a_{12} \; , \\ \ldots & & \\ 1 \cdot u_{1n} & = & a_{1n} \Rightarrow \mathbf{u_{1n}} = a_{1n} \; , \end{array}$$

$$\Rightarrow \quad \mathbf{u_{1j}} = a_{1j}, \quad j = 1, 2, \ldots, n \; .$$

1ª coluna de L: Fazendo o produto de todas as linhas de L, (da 2ª até a nª), pela 1ª coluna de U e igualando com os elementos da 1ª coluna de A, (abaixo da diagonal principal), obtemos ,

$$\begin{array}{rcl} \ell_{21} \; u_{11} & = & a_{21} \Rightarrow \boldsymbol{\ell_{21}} = \dfrac{a_{21}}{u_{11}} \; , \\ \ell_{31} \; u_{11} & = & a_{31} \Rightarrow \boldsymbol{\ell_{31}} = \dfrac{a_{31}}{u_{11}} \; , \\ \ldots & & \\ \ell_{n1} \; u_{11} & = & a_{n1} \Rightarrow \boldsymbol{\ell_{n1}} = \dfrac{a_{n1}}{u_{11}} \; , \end{array}$$

$$\Rightarrow \quad \boldsymbol{\ell_{i1}} = \frac{a_{i1}}{u_{11}}, \quad i = 2, \ldots, n \; .$$

2ª linha de U: Fazendo o produto da 2ª linha de L por todas as colunas de U, (da 2ª até a nª), e igualando com os elementos da 2ª linha de A, (da diagonal principal em diante), obtemos ,

$$\begin{array}{rcl} \ell_{21} \; u_{12} + u_{22} & = & a_{22} \Rightarrow \mathbf{u_{22}} = a_{22} - \ell_{21} u_{12} \; , \\ \ell_{21} \; u_{13} + u_{23} & = & a_{23} \Rightarrow \mathbf{u_{23}} = a_{23} - \ell_{21} u_{13} \; , \\ \ldots & & \\ \ell_{21} \; u_{1n} + u_{2n} & = & a_{2n} \Rightarrow \mathbf{u_{2n}} = a_{2n} - \ell_{21} u_{1n} \; , \end{array}$$

$$\Rightarrow \quad \mathbf{u_{2j}} = a_{2j} - \ell_{21} u_{1j}, \quad j = 3, \ldots, n \; .$$

2ª coluna de L: Fazendo o produto de todas as linhas de L (da 3ª até a nª) pela 2ª coluna de U e igualando com os elementos da 2ª coluna de A, (abaixo da diagonal principal), obtemos ,

$$\begin{array}{rcl} \ell_{31} \; u_{12} + \ell_{32} u_{22} & = & a_{32} \Rightarrow \boldsymbol{\ell_{32}} = \dfrac{a_{32} - \ell_{31} u_{12}}{u_{22}} \; , \\ \ell_{41} \; u_{12} + \ell_{42} u_{22} & = & a_{42} \Rightarrow \boldsymbol{\ell_{42}} = \dfrac{a_{42} - \ell_{41} u_{12}}{u_{22}} \; , \\ \ldots & & \\ \ell_{n1} \; u_{12} + \ell_{n2} u_{22} & = & a_{n2} \Rightarrow \boldsymbol{\ell_{n2}} = \dfrac{a_{n2} - \ell_{n1} u_{12}}{u_{22}} \; , \end{array}$$

$$\Rightarrow \quad \boldsymbol{\ell_{i2}} = \frac{a_{i2} - \ell_{i2} u_{12}}{u_{22}}, \quad i = 3, \ldots, n \; .$$

Se continuarmos calculando 3ª linha de U, 3ª coluna de L, 4ª linha de U, 4ª coluna de L, etc ..., teremos as fórmulas gerais:

$$\begin{cases} \mathbf{u_{ij}} = a_{ij} - \sum_{k=1}^{i-1} \ell_{ik} u_{kj}, & i \leq j, \\ \boldsymbol{\ell_{ij}} = \left(a_{ij} - \sum_{k=1}^{j-1} \ell_{ik} u_{kj} \right) / \; u_{jj}, & i > j. \end{cases} \tag{4.7}$$

Observe que a convenção usual de $\sum_{j=1}^{k} \equiv 0$ se $k < 1$, deve ser utilizada aqui.

Aplicação à Solução de Sistemas Lineares

Vejamos agora como podemos aplicar a decomposição LU para obtermos a solução de sistemas lineares.

Seja o sistema $Ax = b$ de ordem n determinado, onde A satisfaz as condições da decomposição LU. Então o sistema $Ax = b$ pode ser escrito como:

$$LUx \;=\; b\ .$$

Assim transformamos o sistema linear $Ax = b$ no sistema equivalente $LUx = b$ cuja solução é facilmente obtida. De fato: fazendo $Ux = y$, a equação acima reduz-se a $Ly = b$. Resolvendo o sistema triangular inferior $Ly = b$, obtemos o vetor y. Substituindo o valor de y no sistema $Ux = y$ obtemos um sistema triangular superior cuja solução é o vetor x que procuramos.

Assim, a aplicação da decomposição LU na resolução de sistemas lineares requer a solução de dois sistemas triangulares.

Exemplo 4.2 - *Seja:*

$$A \;=\; \begin{pmatrix} 5 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 1 & 3 \end{pmatrix}\ .$$

a) Verificar se A satisfaz as condições da decomposição LU.

b) Decompor A em LU.

c) Através da decomposição LU, calcular o determinante de A.

d) Resolver o sistema $Ax = b$, *onde* $b = (0,\ -7,\ -5)^t$, *usando a decomposição LU.*

Solução:

a) Para que A satisfaça as condições da decomposição LU devemos ter: $det(A_1) \neq 0$ e $det(A_2) \neq 0$. Temos que : $det(A_1) = 5 \neq 0$ e $det(A_2) = -1 \neq 0$. Logo A satisfaz as condições do teorema.

b) Usando as fórmulas (4.7), obtemos:

Para a 1ª linha de U:

$$u_{1j} \;=\; a_{1j}\ , \quad j \;=\; 1,2,3 \;\Rightarrow\; u_{11} = 5\ , \quad u_{12} = 2\ , \quad u_{13} = 1\ .$$

Para a 1ª coluna de L:

$$\ell_{i1} \;=\; \frac{a_{i1}}{u_{11}}\ , \quad i \;=\; 2,3 \;\Rightarrow\; \ell_{21} = \frac{3}{5}\ , \quad \ell_{31} = \frac{1}{5}\ .$$

Para a 2ª linha de U:

$$u_{2j} \;=\; a_{2j} - \ell_{21}\, u_{1j}, \quad j \;=\; 2,3 \;\Rightarrow$$

$$u_{22} \;=\; 1 - \frac{3}{5} \times 2 \;=\; -\frac{1}{5}\ ,\ u_{23} \;=\; 4 - \frac{3}{5} \times 1 \;=\; \frac{17}{5}\ .$$

Para a 2ª coluna de L:

$$\ell_{i2} = \frac{a_{i2} - \ell_{i1}\, u_{12}}{u_{22}}, \quad i = 3 \Rightarrow \ell_{32} = \frac{1 - \frac{1}{5} \times 2}{-\frac{1}{5}} = -3 \, .$$

E finalmente para 3ª linha de U, obtemos:

$$u_{33} = a_{33} - \ell_{31}\, u_{13} - \ell_{32}\, u_{23} \Rightarrow u_{33} = 3 - \frac{1}{5} \times 1 - (-3) \times \frac{17}{5} = 13 \, .$$

Então:

$$L = \begin{pmatrix} 1 & & \bigcirc \\ 3/5 & 1 & \\ 1/5 & -3 & 1 \end{pmatrix} \; ; \quad U = \begin{pmatrix} 5 & 2 & 1 \\ & -1/5 & 17/5 \\ \bigcirc & & 13 \end{pmatrix} .$$

c) $det(A) = u_{11}\, u_{22}\, u_{33} \Rightarrow det(A) = 5 \times \left(-\frac{1}{5}\right) \times 13 \Rightarrow det(A) = -13.$

d) Para obter a solução do sistema $Ax = b$, devemos resolver dois sistemas triangulares: $Ly = b$ e $Ux = y$.

d.1) Assim de $Ly = b$, isto é, de:

$$\begin{pmatrix} 1 & & \bigcirc \\ 3/5 & 1 & \\ 1/5 & -3 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -7 \\ -5 \end{pmatrix} ,$$

obtemos:

$$y_1 = 0 \, ,$$

$$\frac{3}{5}\, y_1 + y_2 = -7 \Rightarrow y_2 = -7 \, ,$$

$$\frac{1}{5}\, y_1 + (-3)\, y_2 + y_3 = -5 \Rightarrow y_3 = -26 \, .$$

Logo, a solução do sistema $Ly = b$ é $y = (0, \; -7, \; -26)^t$.

d.2) De $Ux = y$, isto é, de:

$$\begin{pmatrix} 5 & 2 & 1 \\ & -1/5 & 17/5 \\ \bigcirc & & 13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -7 \\ -26 \end{pmatrix} ,$$

segue que:

$$13\, x_3 = -26 \Rightarrow x_3 = -2 \, ,$$

$$-\frac{1}{5}\, x_2 + \frac{17}{5}\, x_3 = -7 \Rightarrow x_2 = 1 \, ,$$

$$5\, x_1 + 2\, x_2 + x_3 = 0 \Rightarrow x_1 = 0 \, .$$

Logo, a solução do sistema $Ux = y$ é $x = (0, \; 1, \; -2)^t$.

Assim, a solução de $Ax = b$, isto é, de:

$$\begin{pmatrix} 5 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -7 \\ -5 \end{pmatrix} \text{ é } x = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} .$$

Exercícios

4.1 - *Aplicando-se o método da decomposição LU à matriz:*

$$A = \begin{pmatrix} \ldots & \ldots & 3 & \ldots \\ 4 & -1 & 10 & 8 \\ \ldots & -3 & 12 & 11 \\ 0 & -2 & 5 & 10 \end{pmatrix},$$

obteve-se as matrizes:

$$L = \begin{pmatrix} \ldots & 0 & \ldots & \ldots \\ 2 & \ldots & \ldots & \ldots \\ 3 & 0 & \ldots & 0 \\ 0 & \ldots & 1 & \ldots \end{pmatrix}, \quad U = \begin{pmatrix} \ldots & -1 & \ldots & 5 \\ \ldots & 1 & \ldots & -2 \\ \ldots & 0 & 3 & -4 \\ 0 & \ldots & 0 & 10 \end{pmatrix}.$$

Preecher os espaços pontilhados com valores adequados.

4.2 - *Considere o sistema:*

$$\begin{cases} 5\,x_1 + 2\,x_2 + x_3 = -12 \\ -\,x_1 + 4\,x_2 + 2\,x_3 = 20 \\ 2\,x_1 - 3\,x_2 + 10\,x_3 = 3 \end{cases}$$

a) *Resolva-o usando decomposição LU.*

b) *Calcule o determinante de A, usando a decomposição.*

4.3 - *Seja A, $n \times n$, decomponível em LU. Sejam A_i, $i = 1, 2, \ldots, n$, os menores principais de ordem i. Mostre que:*

$$u_{ii} = \frac{\Delta_i}{\Delta_{i-1}}, \quad i = 1, 2, \ldots, n,$$

onde :

$$\Delta_i = detA_i, \quad \Delta_n = detA, \quad e \quad \Delta_0 = 1.$$

4.4 - *Considere a matriz A, $n \times n$, com todas as sub-matrizes principais não singulares. Exiba as fórmulas da decomposição LU, onde L é matriz triangular inferior e U é matriz triangular superior com 1 na diagonal.*

4.5 - *Resolver o sistema $Ax = b$, onde:*

$$A = \begin{pmatrix} 2 & 3 & -1 \\ 1 & 0 & 2 \\ 0 & 3 & -1 \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}; \quad b = \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix},$$

usando a decomposição LU do exercício 4.4.

4.6 - *Mostre que se A satisfaz as hipóteses da decomposição LU então A se decompõe de maneira única no produto LDU, onde L e U são matrizes triangulares inferior e superior, respectivamente, ambas com 1 na diagonal, e D é matriz diagonal. Além disso, $det(A) = d_{11}d_{22}\ldots d_{nn}$.*

4.7 - *Mostre que se A é uma matriz simétrica e satisfaz as hipóteses da decomposição LU então $A = LDU$ implica $U = L^t$ (transposta de L).*

4.8 - *Mostre que se A é uma matriz simétrica, positiva definida e satisfaz as hipóteses da decomposição LU então $A = LDL^t$ onde os elementos diagonais de D são todos positivos.*

4.3 Método de Eliminação de Gauss

Seja o sistema linear $Ax = b$, onde A tem todas as submatrizes principais não singulares.

O **método de Eliminação de Gauss**, também chamado de **método de Gauss Simples**, consiste em transformar o sistema dado num sistema triangular equivalente através de uma sequência de operações elementares sobre as linhas do sistema original, isto é, o sistema equivalente é obtido através da aplicação repetida da operação:

> "*substituir uma equação pela diferença entre essa mesma equação e uma outra equação multiplicada por uma constante diferente de zero*".

É claro que tal operação não altera a solução do sistema, isto é, obtem-se com ela outro sistema equivalente ao original. O objetivo é organizar essa sequência de operações de tal forma que o sistema linear resultante seja triangular superior.

Descrição do algoritmo:

Considere o sistema linear dado por (4.1). Em primeiro lugar montamos a matriz aumentada:

$$\left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3n}^{(1)} & b_3^{(1)} \\ \dots & \dots & & & & \\ a_{n1}^{(1)} & a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right),$$

onde para $i,\ j = 1, 2, \dots, n,\ a_{ij}^{(1)} = a_{ij}$ e $b_i^{(1)} = b_i$.

Por hipótese temos que $a_{11}^{(1)} \neq 0$, pois $det(A_1) \neq 0$.

Primeiro Passo: Eliminar a incógnita x_1 da **2ª, 3ª, ...,** nª equações (isto é, zerar os elementos da primeira coluna abaixo da diagonal); para isso, substituímos a **2ª**, **3ª**, ..., nª equações, respectivamente,

pela diferença entre a **2ª** equação e a **1ª** equação multiplicada por $\dfrac{a_{21}^{(1)}}{a_{11}^{(1)}}$,

pela diferença entre a **3ª** equação e a **1ª** equação multiplicada por $\dfrac{a_{31}^{(1)}}{a_{11}^{(1)}}$,

\- - - - - - - - - - - -

pela diferença entre a nª equação e a **1ª** equação multiplicada por $\dfrac{a_{n1}^{(1)}}{a_{11}^{(1)}}$.

Passamos então da matriz inicial à matriz:

$$\left(\begin{array}{cccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} & & b_1^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} & & b_2^{(2)} \\ & \dots & & & & & \vdots \\ & a_{n2}^{(2)} & a_{n3}^{(2)} & \dots & a_{nn}^{(2)} & & b_n^{(2)} \end{array}\right),$$

onde:

$$\left\{\begin{array}{l} a_{ij}^{(2)} = a_{ij}^{(1)} - a_{1j}^{(1)} \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}} , \\ \qquad\qquad\qquad\qquad i = 2,3,\ldots,n \; ; \\ \qquad\qquad\qquad\qquad j = 1,2,\ldots,n \; . \\ b_i^{(2)} = b_i^{(1)} - b_1^{(1)} \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}} \qquad , \end{array}\right.$$

Observe que da fórmula acima deduzimos que:

$$a_{22}^{(2)} = a_{22}^{(1)} - a_{12}^{(1)} \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{a_{22}^{(1)} a_{11}^{(1)} - a_{12}^{(1)} a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{det(A_2)}{a_{11}^{(1)}} \neq 0 \; ,$$

pois por hipótese $det(A_2) \neq 0$ e $det(A_1) = a_{11}^{(1)} \neq 0$.

Segundo Passo: Eliminar a incógnita x_2 da **3ª, 4ª, ..., nª** equações (isto é, zerar os elementos da segunda coluna abaixo da diagonal); para isso, substituímos a **3ª**, **4ª**, ..., nª equações, respectivamente,

pela diferença entre a **3ª** equação e a **2ª** equação multiplicada por $\dfrac{a_{32}^{(2)}}{a_{22}^{(2)}}$,

pela diferença entre a **4ª** equação e a **2ª** equação multiplicada por $\dfrac{a_{42}^{(2)}}{a_{22}^{(2)}}$,

- - - - - - - - - - - -

pela diferença entre a **nª** equação e a **2ª** equação multiplicada por $\dfrac{a_{n2}^{(2)}}{a_{22}^{(2)}}$.

Obtemos então a matriz:

$$\left(\begin{array}{cccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1n}^{(1)} & & b_1^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2n}^{(2)} & & b_2^{(2)} \\ & & a_{33}^{(3)} & \dots & a_{3n}^{(3)} & & b_3^{(3)} \\ & & \dots & & & & \vdots \\ & & a_{n3}^{(3)} & \dots & a_{nn}^{(3)} & & b_n^{(3)} \end{array}\right),$$

onde:

$$\left\{ \begin{array}{lllll} a_{ij}^{(3)} & = & a_{ij}^{(2)} & - & a_{2j}^{(2)} \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} \, , \\ & & & & \qquad i = 3, 4, \ldots, n \; ; \\ & & & & \qquad j = 2, 3, \ldots, n \; . \\ b_{i}^{(3)} & = & b_{i}^{(2)} & - & b_{2}^{(2)} \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} \, , \end{array} \right.$$

E assim sucessivamente, chegaremos ao **(n-1)º Passo**. Temos que $a_{n-1,n-1}^{(n-1)} \neq 0$, pois por hipótese $det(A_{n-1}) \neq 0$.

(n-1)º Passo: Devemos eliminar a incógnita x_{n-1} da nª equação (isto é, zerar o elemento da $(n-1)$ª coluna abaixo da diagonal); para isso, substituímos a **n**ª equação

pela diferença entre a **n**ª equação e a **(n-1)**ª equação multiplicada por $\dfrac{a_{n,n-1}^{(n-1)}}{a_{n-1,n-1}^{(n-1)}}$, e assim, obtemos a matriz:

$$\left(\begin{array}{cccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1,n-1}^{(1)} & a_{1n}^{(1)} & b_1^{(1)} \\ & a_{22}^{(2)} & a_{23}^{(2)} & \ldots & a_{2,n-1}^{(2)} & a_{2n}^{(2)} & b_2^{(2)} \\ & & a_{33}^{(3)} & \ldots & a_{3,n-1}^{(3)} & a_{3n}^{(3)} & b_3^{(3)} \\ & & & \ldots & \ldots & \ldots & \vdots \\ & & & & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} & b_{n-1}^{(n-1)} \\ & & & & & a_{nn}^{(n)} & b_n^{(n)} \end{array} \right) ,$$

onde:

$$\left\{ \begin{array}{lllll} a_{ij}^{(n)} & = & a_{ij}^{(n-1)} & - & a_{n-1,j}^{(n-1)} \dfrac{a_{i,n-1}^{(n-1)}}{a_{n-1,n-1}^{(n-1)}} \, , \\ & & & & \qquad i = n \; ; \\ & & & & \qquad j = n-1, n \; . \\ b_{i}^{(n)} & = & b_{i}^{(n-1)} & - & b_{n-1}^{(n-1)} \dfrac{a_{i,n-1}^{(n-1)}}{a_{n-1,n-1}^{(n-1)}} \, , \end{array} \right.$$

Assim, de um modo geral, o **(k)º Passo** do método de Eliminação de Gauss, é obtido por:

$$\left\{ \begin{array}{lllll} \mathbf{a_{ij}^{(k+1)}} & = & a_{ij}^{(k)} & - & a_{k,j}^{(k)} \dfrac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \, , \\ & & & & \qquad k = 1, 2, \ldots, n-1 \; , \\ & & & & \qquad i = k+1, \ldots, n \; , \\ & & & & \qquad j = k, k+1, \ldots, n \; . \\ \mathbf{b_{i}^{(k+1)}} & = & b_{i}^{(k)} & - & b_{k}^{(k)} \dfrac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} \, , \end{array} \right. \tag{4.8}$$

Portanto, o sistema triangular obtido,

$$\left\{\begin{array}{rcl}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + a_{13}^{(1)}x_3 + \ldots + a_{1,n-1}^{(1)}x_{n-1} + a_{1n}^{(1)}x_n & = & b_1^{(1)} \\
+ a_{22}^{(2)}x_2 + a_{23}^{(2)}x_3 + \ldots + a_{2,n-1}^{(2)}x_{n-1} + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
a_{33}^{(3)}x_3 + \ldots + a_{3,n-1}^{(3)}x_{n-1} + a_{3n}^{(3)}x_n & = & b_3^{(3)} \\
\ldots\ \ldots & & \\
a_{n-1,n-1}^{(n-1)}x_{n-1} + a_{n-1,n}^{(n-1)}x_n & = & b_{n-1}^{(n-1)} \\
a_{nn}^{(n)}x_n & = & b_n^{(n)}
\end{array}\right.$$

é equivalente ao original.

Observações:

1) No 2º passo, repetimos o processo como se não existisse a 1ª linha e a 1ª coluna da 2ª matriz, isto é, todas as operações são realizadas em função da 2ª linha da matriz obtida no 1º passo. De um modo geral, no $(k)^o$ passo, repetimos o processo como se não existissem as $(k-1)$ primeiras linhas e as $(k-1)$ primeiras colunas da k^a matriz, isto é, todas as operações são realizadas em função da linha (k) da matriz obtida no passo $(k-1)$.

2) A operação: substituir uma equação pela diferença entre essa mesma equação e uma outra equação multiplicada por uma constante diferente de zero, é equivalente a realizarmos o seguinte cálculo (descreveremos para o (k^o) passo):

2.1) Determinar as constantes $a_{ik}^{(k)}/a_{kk}^{(k)}$,

2.2) Um elemento $a_{ij}^{(k+1)}$ será então obtido fazendo-se na matriz anterior a diferença entre o elemento que ocupa a mesma posição, isto é, $(a_{ij}^{(k)})$ e o produto da constante $(a_{ik}^{(k)}/a_{kk}^{(k)})$ pelo elemento que se encontra na mesma coluna da linha k, ou seja, pelo elemento $(a_{kj}^{(k)})$.

3) O elemento $a_{kk}^{(k)}$ é o pivô do kº passo.

Exemplo 4.3 - *Resolver o sistema:*

$$\begin{pmatrix} 6 & 2 & -1 \\ 2 & 4 & 1 \\ 3 & 2 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 7 \\ 13 \end{pmatrix},$$

usando o método de Eliminação de Gauss.

Solução: Montamos incialmente a matriz, 3×4:

$$\left(\begin{array}{rrr|r} 6 & 2 & -1 & 7 \\ 2 & 4 & 1 & 7 \\ 3 & 2 & 8 & 13 \end{array}\right).$$

1º Passo: Temos que:

$$\frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{2}{6} = \frac{1}{3}\,, \quad \text{e} \quad \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{3}{6} = \frac{1}{2}\,.$$

Assim:

$$a_{21}^{(2)} = a_{21}^{(1)} - a_{11}^{(1)}\,\frac{a_{21}^{(1)}}{a_{11}^{(1)}} \Rightarrow a_{21}^{(2)} = 2 - 6 \times \frac{1}{3} \Rightarrow a_{21}^{(2)} = 0\,,$$

$$a_{22}^{(2)} = a_{22}^{(1)} - a_{12}^{(1)}\,\frac{a_{21}^{(1)}}{a_{11}^{(1)}} \Rightarrow a_{22}^{(2)} = 4 - 2 \times \frac{1}{3} \Rightarrow a_{22}^{(2)} = \frac{10}{3}\,,$$

$$a_{23}^{(2)} = a_{23}^{(1)} - a_{13}^{(1)}\,\frac{a_{21}^{(1)}}{a_{11}^{(1)}} \Rightarrow a_{23}^{(2)} = 1 - (-1) \times \frac{1}{3} \Rightarrow a_{23}^{(2)} = \frac{4}{3}\,,$$

$$b_{2}^{(2)} = b_{2}^{(1)} - b_{1}^{(1)}\,\frac{a_{21}^{(1)}}{a_{11}^{(1)}} \Rightarrow b_{2}^{(2)} = 7 - 7 \times \frac{1}{3} \Rightarrow b_{2}^{(2)} = \frac{14}{3}\,,$$

$$a_{31}^{(2)} = a_{31}^{(1)} - a_{11}^{(1)}\,\frac{a_{31}^{(1)}}{a_{11}^{(1)}} \Rightarrow a_{31}^{(2)} = 3 - 6 \times \frac{1}{2} \Rightarrow a_{31}^{(2)} = 0\,,$$

$$a_{32}^{(2)} = a_{32}^{(1)} - a_{12}^{(1)}\,\frac{a_{31}^{(1)}}{a_{11}^{(1)}} \Rightarrow a_{32}^{(2)} = 2 - 2 \times \frac{1}{2} \Rightarrow a_{32}^{(2)} = 1\,,$$

$$a_{33}^{(2)} = a_{33}^{(1)} - a_{13}^{(1)}\,\frac{a_{31}^{(2)}}{a_{11}^{(1)}} \Rightarrow a_{33}^{(2)} = 8 - (-1) \times \frac{1}{2} \Rightarrow a_{33}^{(2)} = \frac{17}{2}\,,$$

$$b_{3}^{(2)} = b_{3}^{(1)} - b_{1}^{(1)}\,\frac{a_{31}^{(1)}}{a_{11}^{(1)}} \Rightarrow b_{3}^{(2)} = 13 - 7 \times \frac{1}{2} \Rightarrow b_{3}^{(2)} = \frac{19}{2}\,.$$

Assim ,obtemos a matriz:

$$\begin{pmatrix} 6 & 2 & -1 & | & 7 \\ 0 & 10/3 & 4/3 & | & 14/3 \\ 0 & 1 & 17/2 & | & 19/2 \end{pmatrix}.$$

2º Passo: Temos que:

$$\frac{a_{32}^{(2)}}{a_{22}^{(2)}} = \frac{1}{10/3} = \frac{3}{10}\,.$$

Assim:

$$a_{32}^{(3)} = a_{32}^{(2)} - a_{22}^{(2)}\,\frac{a_{32}^{(2)}}{a_{22}^{(2)}} \Rightarrow a_{32}^{(3)} = 1 - \frac{10}{3} \times \frac{3}{10} \Rightarrow a_{32}^{(3)} = 0\,,$$

$$a_{33}^{(3)} = a_{33}^{(2)} - a_{23}^{(2)}\,\frac{a_{32}^{(2)}}{a_{22}^{(2)}} \Rightarrow a_{33}^{(3)} = \frac{17}{2} - \frac{4}{3} \times \frac{3}{10} \Rightarrow a_{33}^{(3)} = \frac{81}{10}\,,$$

$$b_{3}^{(3)} = b_{3}^{(1)} - a_{2}^{(2)}\,\frac{a_{32}^{(2)}}{a_{22}^{(2)}} \Rightarrow b_{3}^{(3)} = \frac{19}{2} - \frac{14}{3} \times \frac{3}{10} \Rightarrow b_{3}^{(3)} = \frac{81}{10}\,.$$

Portanto, obtemos:

$$\left(\begin{array}{ccccc} 6 & 2 & -1 & | & 7 \\ & 10/3 & 4/3 & | & 14/3 \\ \bigcirc & & 81/10 & | & 81/10 \end{array}\right),$$

ou seja:

$$\left(\begin{array}{ccc} 6 & 2 & -1 \\ 0 & 10/3 & 4/3 \\ 0 & 0 & 81/10 \end{array}\right) \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 7 \\ 14/3 \\ 81/10 \end{array}\right).$$

Resolvendo então o sitema linear, obtemos:

$$\frac{81}{10}\, x_3 = \frac{81}{10} \Rightarrow x_3 = 1\,,$$

$$\frac{10}{3}\, x_2 + \frac{4}{3}\, x_3 = \frac{14}{3} \Rightarrow x_2 = 1\,,$$

$$6\, x_1 + 2\, x_2 - x_3 = 7 \Rightarrow x_1 = 1\,.$$

Assim, a solução de:

$$\left(\begin{array}{ccc} 6 & 2 & -1 \\ 3 & 4 & 1 \\ 3 & 2 & 8 \end{array}\right) \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 7 \\ 7 \\ 13 \end{array}\right) \text{ é } x = \left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right).$$

Observação: Se em algum passo k encontrarmos $a_{kk}^{(k)} = 0$, isso significa que $det(A_k) = 0$. Nesse caso, o sistema ainda pode ter solução determinada (basta que $det(A) \neq 0$). O método pode ser continuado simplesmente permutando a k^{a} equação com qualquer outra abaixo cujo coeficiente da k^{a} incógnita seja $\neq 0$.

Exemplo 4.4 - *Resolver, usando o método de Eliminação de Gauss, o sistema:*

$$\left\{\begin{array}{rcrcrcr} 3x_1 & + & 3x_2 & + & x_3 & = & 7 \\ 2x_1 & + & 2x_2 & - & x_3 & = & 3 \\ x_1 & - & x_2 & + & 5x_3 & = & 5 \end{array}\right.$$

Solução: Aplicando o método de Eliminação de Gauss à matriz:

$$\left(\begin{array}{ccccc} 3 & 3 & 1 & | & 7 \\ 2 & 2 & -1 & | & 3 \\ 1 & -1 & 5 & | & 5 \end{array}\right),$$

obtemos:

$$\left(\begin{array}{ccccc} 3 & 3 & 1 & | & 7 \\ 0 & 0 & -5/3 & | & -5/3 \\ 0 & -2 & 14/3 & | & 8/3 \end{array}\right).$$

Vemos aqui que o elemento $a_{22}^{(2)} = 0$ e como já dissemos anteriormente, isso significa que $det(A_2) = 0$. De fato: $det(A_2) = \left|\begin{array}{cc} 3 & 3 \\ 2 & 2 \end{array}\right| = 0$.

Como o elemento $a_{32}^{(2)} \neq 0$, permutamos a 3ª equação com a 2ª e assim obtemos a matriz:

$$\begin{pmatrix} 3 & 3 & 1 & | & 7 \\ & -2 & 14/3 & | & 8/3 \\ \bigcirc & & -5/3 & | & -5/3 \end{pmatrix} .$$

a qual já está na forma triangular. Assim a solução de:

$$\begin{cases} 3x_1 + 3x_2 + x_3 = 7 \\ 2x_1 + 2x_2 - x_3 = 3 \\ x_1 - x_2 + 5x_3 = 5 \end{cases} \quad é \ x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} .$$

Observações:

1) O método de Eliminação de Gauss pode ser interpretado como um método para obtenção das matrizes L e U da decomposição LU. De fato, chamando de $(A|b)^{(1)}$ a matriz aumentada, o cálculo feito para obtenção de $(A|b)^{(2)}$, é equivalente a multiplicar $(A|b)^{(1)}$ por uma matriz M_1, onde:

$$M_1 = \begin{pmatrix} 1 & & & \\ -m_{21} & 1 & & \\ \vdots & & \ddots & \\ -m_{n1} & & & 1 \end{pmatrix} , \quad \text{com} \quad m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \quad i = 2, 3 \ldots , n .$$

Assim, $(A|b)^{(2)} = M_1(A|b)^{(1)}$.

De maneira semelhante: $(A|b)^{(3)} = M_2(A|b)^{(1)}$, onde:

$$M_2 = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & -m_{32} & 1 & & \\ & \vdots & & \ddots & \\ & -m_{n2} & & & 1 \end{pmatrix} , \quad \text{com} \quad m_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} \quad i = 3, \ldots , n ,$$

e assim sucessivamente. Logo:

$$(A|b)^{(n)} = M_{n-1}(A|b)^{(n-1)} = \ldots = \underbrace{M_{n-1} \ldots M_1}_{M}(A|b)^{(1)} .$$

Assim, temos: $A^{(n)} = MA^{(1)} = MA = U$ (onde U é matriz triangular superior da Decomposição LU). Como M é um produto de matrizes não singulares então é inversível; logo existe $M^{-1} = M_1^{-1} M_2^{-1} \ldots M_{n-1}^{-1}$ e portanto $A = M^{-1}U$.

É fácil verificar que:

$$M^{-1} = \begin{pmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ & \vdots & & \ddots & \\ m_{n1} & m_{n2} & & & 1 \end{pmatrix} = L ,$$

onde L é a matriz triangular inferior da decomposição LU.

Assim o método de Eliminação de Gauss nada mais é do que o método da Decomposição LU.

Observe que devemos resolver apenas o sistema $Ux = b^{(n)}$, desde que o vetor final $b^{(n)}$ é obtido de b através da equação: $b = Lb^{(n)}$. Assim, se $Ax = b$, $A = LU$, $b = Lb^{(n)}$, obtemos:

$$LUx = Lb^{(n)} \Rightarrow Ux = b^{(n)} ,$$

pois, como já dissemos, L é não singular. Portanto, o vetor solução x é obtido resolvendo-se apenas um sistema triangular.

2) O exemplo 4.4 apresenta um sistema de equações para o qual as hipóteses do Teorema 4.1 não são satisfeitas. Entretanto, resolvemos o sistema utilizando a estratégia de troca de linhas. Para visualizar porque isso foi possível, observe que podemos construir uma matriz P, chamada **matriz de Permutação**, a qual será formada pela permutação das linhas da matriz identidade, de forma que em cada linha o único elemento não nulo é igual a 1. Se durante o processo não permutamos as linhas do sistema então P é a matriz identidade. Agora se durante o processo permutamos a linha i com a linha j então na matriz P a linha i será permutada com a linha j. No exemplo 4.4, permutamos durante o processo a linha 2 com a linha 3. Assim:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 0 & 1 \end{pmatrix} , \quad U = \begin{pmatrix} 3 & 3 & 1 \\ 0 & -2 & 14/3 \\ 0 & 0 & -5/3 \end{pmatrix} .$$

e a matriz P nesse caso será:

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} .$$

Pode ser visto facilmente que com a troca de linhas a decomposição obtida não satisfaz a igualdade $A = LU$, mas satisfaz $PA = LU$, isto é, o produto LU reproduz a matriz A com suas linhas permutadas. Assim foi possível resolver o sistema do exemplo 4.4 pois a troca de linhas na decomposição funciona como se tivéssemos efetuado troca de linhas no sistema antes de começarmos a decomposição. A estratégia de troca de linhas para resolver um sistema é extremamente útil para os métodos baseados na decomposição LU e é conhecida como pivotamento. Descreveremos mais adiante o método de Eliminação de Gauss com pivotamento parcial.

Exercícios

4.9 - *Considere o sistema:*

$$\begin{pmatrix} 2 & -3 & 1 \\ 4 & -6 & -1 \\ 1 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -5 \\ -7 \\ 4 \end{pmatrix}$$

a) *Resolva-o pelo método de Eliminação de Gauss,*

b) *Calcule o determinante de A, usando a matriz triangular obtida no item* **a)**.

4.10 - *Verificar, usando o método de Eliminação de Gauss, que o sistema :*

$$\begin{cases} x_1 + 2x_2 + x_3 = 3 \\ 2x_1 + 3x_2 + x_3 = 5 \\ 3x_1 + 5x_2 + 2x_3 = 1 \end{cases}$$

não tem solução.

4.11 - *Usando o método de Eliminação de Gauss, verificar que o sistema:*

$$\begin{cases} x_1 & + & 4\,x_2 & + & \alpha\,x_3 & = & 6 \\ 2\,x_1 & - & x_2 & + & 2\,\alpha x_3 & = & 3 \\ \alpha\,x_1 & + & 3\,x_2 & + & x_3 & = & 5 \end{cases}$$

a) *possui uma única solução quando* $\alpha = 0$*;*

b) *infinitas soluções quando* $\alpha = 1$ *e*

c) *não tem solução quando* $\alpha = -1$.

4.4 Método de Gauss-Compacto

Como vimos, o método de Eliminação da Gauss nada mais é do que o método da Decomposição LU: a matriz triangular superior obtida ao final da aplicação desse método é a matriz U da decomposição LU e a matriz L é a matriz formada pelos multiplicadores (as constantes $\frac{a_{ik}^{(k)}}{a_{kk}^{k}}$ do $k^{\underline{o}}$ passo).

Descreveremos agora uma maneira prática de se obter as matrizes L e U, bem como armazená-las de uma forma compacta. Tal método recebe o nome de **método de Gauss-Compacto**. A vantagem desse método é a de economizar espaço na memória, pois ambas as matrizes L e U são armazenadas sobre a matriz original A. O único incoveniente é que a matriz A é destruída. O termo independente b, é transformado juntamente com a matriz A, como no método de Gauss simples. Esse procedimento é bastante usual e conveniente para simplificar a programação desses métodos. Ao final teremos armazenados as matrizes L, U e o termo independente modificado. A solução do sistema será obtida resolvendo-se apenas um sistema triangular superior.

Descrição do algoritmo:

Considere o sistema linear, de ordem n, dado por (4.1). Em primeiro lugar montamos a matriz, $n \times n+1$:

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & | & a_{1,n+1} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & | & a_{2,n+1} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} & | & a_{3,n+1} \\ \dots & \dots & & & & | & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & | & a_{n,n+1} \end{pmatrix},$$

onde $a_{i,n+1} \; = \; b_i, \; i = 1, 2, \dots, n$.

A seguir construímos a matriz, $n \times n+1$, onde os termos independentes $b_i = a_{i,n+1}, \; i = 1, \dots, n$ por serem obtidos da mesma maneira que os elementos u_{ij} serão chamados $u_{i,n+1}, \; i = 1, \dots, n$. Assim, sobre a matriz original armazenamos a matriz:

$$\begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} & | & u_{1,n+1} \\ \ell_{21} & u_{22} & u_{23} & \dots & u_{2n} & | & u_{2,n+1} \\ \ell_{31} & \ell_{32} & u_{33} & \dots & u_{3n} & | & u_{3,n+1} \\ \dots & \dots & & & & | & \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \dots & u_{nn} & | & u_{n,n+1} \end{pmatrix},$$

a qual é obtida através das fórmulas da decomposição LU, (fórmulas (4.7)), levando em consideração a lei de formação das mesmas, na seguinte ordem:

a) **1ª linha** : $u_{1j} = a_{1j},\ j = 1, 2, \ldots, n+1,$

(isto significa que 1ª linha é igual a 1ª linha da matriz original) .

b) **1ª coluna:** $\ell_{i1} = \frac{a_{i1}}{u_{11}}\ ;\ i = 2, \ldots, n;$

(isto significa que a 1ª coluna é igual ao elemento que ocupa a mesma posição na matriz original dividido por a_{11} pois $u_{11} = a_{11}$).

c) **2ª linha** : $u_{2j} = a_{2j} - l_{21}u_{1j},\ j = 2, \ldots, n+1,$

(isto significa que a 2ª linha é igual a diferença entre o elemento que ocupa a mesma posição na matriz original e o produto da 2ª linha pela coluna j na segunda matriz , com $j = 2, \ldots, n+1$) .

d) **2ª coluna:** $\ell_{i2} = \frac{a_{i2} - l_{i1}u_{12}}{u_{22}}\ ;\ i = 3, \ldots, n;$

(isto significa, que a 2ª coluna é igual a diferença entre o elemento que ocupa a mesma posição na matriz original e o produto da linha i pela 2ª coluna, tudo dividido por u_{22}, com $i = 3, \ldots, n$).

e) Segue-se calculando: 3ª linha, 3ª coluna,. . .,

lembrando que as linhas são calculadas da diagonal (inclusive) em diante, e as colunas da diagonal (exclusive) para baixo. Assim, o elemento $\mathbf{u_{ij}}$ é a diferença entre o elemento que ocupa a mesma posição na matriz original e a soma do produto ordenado da linha i pela coluna j (2ª matriz), até a linha $i-1$. Para o elemento $\ell_{\mathbf{ij}}$ o procedimento é análogo (limitado pela coluna $j-1$) e dividido pelo elemento diagonal na coluna.

Observações:

1) Produto ordenado significando: 1º elemento da linha i multiplicado pelo 1º elemento da coluna j, 2º× 2º,. . ..

2) Calculadas todas as linhas e colunas resolve-se o sistema $Ux = b'$, onde U está indicada na 2ª matriz e b' é a última coluna da 2ª matriz.

3) Quem não consegue visualizar a maneira prática para montar a 2ª matriz deve recorrer às formulas (4.7), para resolver o sistema pelo método de Gauss-Compacto.

4) Dado vários sistemas associados a uma mesma matriz, podemos resolvê-los de uma só vez pelo método da Gauss-Compacto. Tais sistemas são chamados de **sistemas matriciais**.

Exemplo 4.5 - *Usando o Método de Gauss-Compacto resolver o sistema matricial:*

$$\begin{pmatrix} 5 & 2 & -1 \\ 3 & 1 & 4 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 & | & y_1 \\ x_2 & | & y_2 \\ x_3 & | & y_3 \end{pmatrix} = \begin{pmatrix} 0 & | & 6 \\ -7 & | & 7 \\ -5 & | & 4 \end{pmatrix} .$$

Solução: Montamos a matriz, 3×5:

$$\begin{pmatrix} 5 & 2 & 1 & | & 0 & | & 6 \\ 3 & 1 & 4 & | & -7 & | & 7 \\ 1 & 1 & 3 & | & -5 & | & 4 \end{pmatrix} \sim \begin{pmatrix} 5 & 2 & 1 & | & 0 & | & 6 \\ 3/5 & -1/5 & 17/5 & | & -7 & | & 17/5 \\ 1/5 & -3 & 13 & | & -26 & | & 13 \end{pmatrix} ;$$

onde, obtivemos os elementos da seguinte maneira:

1ª linha: (igual a 1ª linha da matriz original). Assim:

$$u_{11} = 5, \quad u_{12} = 2, \quad u_{13} = 1, \quad u_{14} = 0, \quad u_{15} = 6 \ .$$

1ª coluna: (igual ao elemento que ocupa a mesma posição na matriz original dividido por a_{11} desde que $u_{11} = a_{11}$). Assim:

$$\ell_{21} = \frac{3}{5}, \quad \ell_{31} = \frac{1}{5} \ .$$

2ª linha: (igual a diferença entre o elemento que ocupa a mesma posição na matriz original e o produto da linha 2 pela coluna j, limitado pela linha 1, na 2ª matriz, Assim:

$$\begin{aligned}
u_{22} &= 1 - \frac{3}{5} \times 2 \ \Rightarrow \ u_{22} = -\frac{1}{5} \ , \\
u_{23} &= 4 - \frac{3}{5} \times 1 \ \Rightarrow \ u_{23} = \frac{17}{5} \ , \\
u_{24} &= -7 - \frac{3}{5} \times 0 \ \Rightarrow \ u_{24} = -7 \ , \\
u_{25} &= 7 - \frac{3}{5} \times 6 \ \Rightarrow \ u_{25} = \frac{17}{5} \ .
\end{aligned}$$

2ª coluna: (igual a diferença entre o elemento que ocupa a mesma posição na matriz original e o produto da linha 3 pela coluna 2, limitado pela coluna 1, dividido pelo elemento da diagonal principal na 2ª matriz). Assim:

$$\ell_{32} = \frac{1 - \frac{1}{5} \times 2}{-\frac{1}{5}} \ \Rightarrow \ \ell_{32} = -3 \ .$$

3ª linha: (igual a diferença entre o elemento que ocupa a mesma posição na matriz original e o produto da linha 3 pela coluna j, limitado pela linha 2, na 2ª matriz). Assim:

$$\begin{aligned}
u_{33} &= 3 - \frac{1}{5} \times 1 - (-3) \times \frac{17}{5} \ \Rightarrow \ u_{33} = 13 \ , \\
u_{34} &= -5 - \frac{1}{5} \times 0 - (-3) \times (-7) \ \Rightarrow \ u_{34} = -26 \ , \\
u_{35} &= 4 - \frac{1}{5} \times 6 - (-3) \times \frac{17}{5} \ \Rightarrow \ u_{35} = 13 \ .
\end{aligned}$$

Assim, resolvendo os sistemas:

$$\text{a)} \begin{pmatrix} 5 & 2 & 1 \\ & -1/5 & 17/5 \\ \bigcirc & & 13 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ -7 \\ -26 \end{pmatrix} ; \text{ obtemos } x = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix} ,$$

e

$$\text{b)} \begin{pmatrix} 5 & 2 & 1 \\ & -1/5 & 17/5 \\ \bigcirc & & 13 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 17/5 \\ 13 \end{pmatrix} ; \text{ obtemos } y = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} .$$

Portanto a solução do sistema matricial:

$$\begin{pmatrix} 5 & 2 & 1 \\ 3 & 1 & 4 \\ 1 & 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 & | & y_1 \\ x_2 & | & y_2 \\ x_3 & | & y_3 \end{pmatrix} = \begin{pmatrix} 0 & | & 6 \\ -7 & | & 7 \\ -5 & | & 4 \end{pmatrix} \text{ é } (x|y) = \begin{pmatrix} 0 & | & 1 \\ 1 & | & 0 \\ -2 & | & 1 \end{pmatrix} .$$

Exercícios

4.12 - *Aplicando-se o método de Gauss-Compacto a um sistema $Ax = b$, foi obtido o esquema:*

$$\begin{pmatrix} \dots & 2 & 1 & \dots & | & 5 \\ 6 & 1 & 0 & 3 & | & 10 \\ \dots & -3 & -5 & 7 & | & 2 \\ 9 & 0 & -2 & -1 & | & 6 \end{pmatrix} \sim \begin{pmatrix} 3 & 2 & 1 & -1 & | & 5 \\ 2 & \dots & -2 & 5 & | & 0 \\ 1 & 5/3 & -8/3 & \dots & | & -3 \\ \dots & 2 & \dots & -63/8 & | & \dots \end{pmatrix}.$$

a) *Preencha os espaços pontilhados com valores adequados.*

b) *Se $x = (x_1,\ x_2,\ x_3,\ x_4)^t$, calcule x_3 e x_4.*

4.13 - *Usando o método de Gauss-Compacto resolver os seguintes sistemas:*

$$(I)\left\{\begin{array}{rcrcrcr} 10\,x_1 & + & x_2 & - & x_3 & = & 10 \\ x_1 & + & 10\,x_2 & + & x_3 & = & 12 \\ 2\,x_1 & - & x_2 & + & 10\,x_3 & = & 11 \end{array}\right.$$

$$(II)\left\{\begin{array}{rcrcrcr} 4\,x_1 & - & 6\,x_2 & - & x_3 & = & -7 \\ 2\,x_1 & - & 3\,x_2 & + & x_3 & = & -5 \\ x_1 & + & 2\,x_2 & + & x_3 & = & 4 \end{array}\right.$$

4.14 - *Resolver o sistema matricial:*

$$\begin{pmatrix} 2 & -1 & 3 \\ 4 & 1 & 2 \\ 1 & 0 & 10 \end{pmatrix} \begin{pmatrix} x_1 & | & y_1 & | & z_1 \\ x_2 & | & y_2 & | & z_2 \\ x_3 & | & y_3 & | & z_3 \end{pmatrix} = \begin{pmatrix} -4 & | & 2 & | & 4 \\ -7 & | & 6 & | & 6 \\ -11 & | & 2 & | & 20 \end{pmatrix},$$

usando o método de Gauss-Compacto.

4.5 Método de Cholesky

No caso em que a matriz do sistema linear é simétrica podemos simplificar os cálculos da decomposição LU significativamente, levando em conta a simetria. Esta é a estratégia do **método de Cholesky**, o qual se baseia no seguinte corolário.

Corolário 4.1 - *Se A é simétrica, positiva definida, então A pode ser decomposta unicamente no produto GG^t, onde G é matriz triangular inferior com elementos diagonais positivos.*

Prova: A prova é imediata a partir do exercício 4.8.

Observe que essa decomposição é possível se a matriz A além de simétrica for também positiva definida (veja definição 4.4).

Esquema Prático para a Decomposição GG^t

Do mesmo modo que na da decomposição LU para obtermos a matriz G, aplicamos a definição de produto e igualdade de matrizes. Seja então:

$$GG^t = \begin{pmatrix} g_{11} & & & & \bigcirc \\ g_{21} & g_{22} & & & \\ g_{31} & g_{32} & g_{33} & & \\ \dots & \dots & \dots & \ddots & \\ g_{n1} & g_{n2} & g_{n3} & \cdots & g_{nn} \end{pmatrix} \begin{pmatrix} g_{11} & g_{21} & g_{31} & \cdots & g_{n1} \\ & g_{22} & g_{32} & \cdots & g_{n2} \\ & & g_{33} & \cdots & g_{n3} \\ & & & \ddots & \vdots \\ \bigcirc & & & & g_{nn} \end{pmatrix}, \text{ e}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \dots & \dots & \dots & & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix}.$$

Desde que existe uma lei de formação para os elementos diagonais e outra para os não diagonais de G, veremos em separado como obter tais fórmulas.

a) **Elementos diagonais de** G .

Os elementos diagonais a_{ii} de A são iguais ao produto da linha i de G pela coluna i de G^t. Veja que este produto é equivalente a multiplicarmos a linha i de G por ela mesma. Portanto:

$$\begin{aligned} a_{11} &= g_{11}^2 , \\ a_{22} &= g_{21}^2 + g_{22}^2 , \\ \dots \\ a_{nn} &= g_{n1}^2 g_{n2}^2 \ + \ \dots \ + \ g_{nn}^2 . \end{aligned}$$

Logo, os elementos diagonais de G são dados por:

$$\begin{cases} \mathbf{g_{11}} = \sqrt{a_{11}} , \\ \mathbf{g_{ii}} = \left(a_{ii} - \sum_{k=1}^{i-1} g_{ik}^2 \right)^{1/2} , \quad i = 2, 3, \dots, n. \end{cases} \tag{4.9}$$

b) **Elementos não diagonais de** G**.**

b.1) **1ª coluna :** Os elementos da 1ª coluna de G, são obtidos igualando-se os elementos da 1ª coluna de A com o produto de cada linha de G pela 1ª coluna de G^t. Observe que este produto pode ser obtido multiplicando-se cada elemento da 1ª coluna de G (abaixo da diagonal) pela 1ª linha de G . Assim:

$$\begin{aligned} a_{21} &= \mathbf{g_{21}} \, g_{11} , \\ a_{31} &= \mathbf{g_{31}} \, g_{11} , \\ \dots \\ a_{n1} &= \mathbf{g_{n1}} \, g_{11} . \end{aligned}$$

$$\Rightarrow \mathbf{g_{i1}} = \frac{a_{i1}}{g_{11}}, \quad i = 2, 3, \dots, n.$$

b.2) **2ª coluna:** Os elementos da 2ª coluna de G, são obtidos igualando-se os elementos da 2ª coluna de A (abaixo da diagonal principal) com o produto de cada linha de G pela 2ª coluna de G^t. Observe que este produto pode ser obtido multiplicando-se cada linha de G (abaixo da diagonal) pela 2ª

linha de G. Assim:

$$\begin{array}{rcl} a_{32} & = & g_{31}\ g_{21}\ +\ \mathbf{g_{32}}\ g_{22}\ , \\ a_{42} & = & g_{41}\ g_{21}\ +\ \mathbf{g_{42}}\ g_{22}\ , \\ \dots & & \\ a_{n2} & = & g_{n1}\ g_{21}\ +\ \mathbf{g_{n2}}\ g_{22}\ , \\ & & \\ \Rightarrow \mathbf{g_{i2}} & = & \dfrac{a_{i2} - g_{i1}g_{21}}{g_{22}}, \quad i = 3, 4, \dots, n. \end{array}$$

Se continuarmos calculando 3ª, 4ª colunas de G, etc ..., teremos a fórmula geral:

$$\left\{ \begin{array}{lcl} \mathbf{g_{i1}} & = & \frac{a_{i1}}{g_{11}}, \quad i = 2, 3, \dots, n, \\ & & \\ \mathbf{g_{ij}} & = & \left(a_{ij}\ -\ \sum_{k=1}^{j-1}\ g_{ik}\ g_{jk} \right)\ /\ g_{jj}\ \ ,\ \ 2 \leq j < i. \end{array} \right. \tag{4.10}$$

Utilizadas numa ordem conveniente as fórmulas (4.9) e (4.10) determinam os elementos da matriz G. Uma ordem conveniente pode ser:

$$g_{11},\ g_{21},\ g_{31}, \dots,\ g_{n1};\ g_{22},\ g_{32}, \dots,\ g_{n2}; \dots, g_{nn}.$$

Isto corresponde a calcularmos os elementos da matriz G por coluna.

Observações:

i) Se A satisfaz as condições do método de Cholesky, a aplicação do método requer menos cálculos que a decomposição LU.

ii) O fato de A ser positiva definida garante que na decomposição teremos somente raízes quadradas de números positivos.

iii) O método de Cholesky pode também ser aplicado a matrizes simétricas que não sejam positivas definidas desde que trabalhemos com aritmética complexa. Entretanto, só usaremos o método de Cholesky se pudermos trabalhar com aritmética real.

iv) Vimos no caso da decomposição LU, que $det(A) = u_{11}u_{22}\dots u_{nn}$, uma vez que os elementos diagonais de L eram unitários. No caso do Método de Cholesky temos que: $A = GG^t$ e portanto:

$$det(A)\ =\ (det\ G)^2\ =\ (g_{11}\ g_{22}\ \dots\ g_{nn})^2$$

Aplicação à Solução de Sistemas Lineares

Vejamos agora como podemos aplicar a decomposição GG^t para obtermos a solução de sistemas lineares.

Seja o sistema $Ax = b$ de ordem n determinado, onde A satisfaz as condições do processo de Cholesky. Uma vez calculado a matriz G a solução de $Ax = b$ fica reduzida, como no método da Decomposição LU, à solução do par de sistemas triangulares:

$$\left\{ \begin{array}{lcl} Gy & = & b\ , \\ & & \\ G^t x & = & y\ . \end{array} \right.$$

Exemplo 4.6 - *Seja:*

$$A = \begin{pmatrix} 4 & 2 & -4 \\ 2 & 10 & 4 \\ -4 & 4 & 9 \end{pmatrix} .$$

a) Verificar se A satisfaz as condições do método de Cholesky.

b) Decompor A em GG^t.

c) Calcular o determinante de A, usando a decomposição obtida .

d) Resolver o sistema $Ax = b$, onde $b = (0,\ 6,\ 5)^t$.

Solução:

a) A matriz A é simétrica. Devemos verificar se é positiva definida. Temos:

$$det(A_1) = 4 > 0 \ , \quad det(A_2) = 36 > 0 \ , \quad det(A_3) = det(A) = 36 > 0 \ .$$

Logo, A satifaz as condições da decomposição GG^t.

b) Usando as fórmulas (4.9) e (4.10), obtemos:

$$\begin{aligned}
g_{11} &= \sqrt{a_{11}} \Rightarrow g_{11} = \sqrt{4} \Rightarrow g_{11} = 2 \ , \\
g_{21} &= \frac{a_{21}}{g_{11}} \Rightarrow g_{21} = \frac{2}{2} \Rightarrow g_{21} = 1 \ , \\
g_{31} &= \frac{a_{31}}{g_{11}} \Rightarrow g_{31} = \frac{-4}{2} \Rightarrow g_{31} = -2 \ , \\
g_{22} &= \left(a_{22} - g_{21}^2\right)^{1/2} \Rightarrow g_{22} = \left(10 - 1^2\right)^{1/2} \Rightarrow g_{22} = 3 \ , \\
g_{32} &= \frac{a_{32} - g_{31} g_{21}}{g_{22}} \Rightarrow g_{32} = \frac{4 - (-2)(1)}{3} \Rightarrow g_{32} = 2, \\
g_{33} &= \left(a_{33} - g_{31}^2 - g_{32}^2\right)^{1/2} \Rightarrow g_{33} = \left(9 - (-2)^2 - 2^2\right)^{1/2} \Rightarrow g_{33} = 1.
\end{aligned}$$

Obtemos então:

$$\begin{pmatrix} 4 & 2 & -4 \\ 2 & 10 & 4 \\ -4 & 4 & 9 \end{pmatrix} = \begin{pmatrix} 2 & & \bigcirc \\ 1 & 3 & \\ -2 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 \\ & 3 & 2 \\ \bigcirc & & 1 \end{pmatrix} .$$

c) $det(A) = \left(g_{11}\ g_{22}\ g_{33}\right)^2 = (2 \times 3 \times 1)^2 = 36.$

d) Para obter a solução do sistema $Ax = b$, devemos resolver dois sistemas triangulares: $Gy = b$ e $G^t x = y$.

d.1) Assim de $Gy = b$, isto é, de:

$$\begin{pmatrix} 2 & & \bigcirc \\ 1 & 3 & \\ -2 & 2 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 6 \\ 5 \end{pmatrix} ,$$

obtemos:

$$2\,y_1 = 0 \Rightarrow y_1 = 0 ,$$

$$y_1 + 3\,y_2 = 6 \Rightarrow y_2 = 2 ,$$

$$-2\,y_1 + 2\,y_2 + y_3 = 5 \Rightarrow y_3 = 1 .$$

Logo a solução do sistema $Gy = b$ é $y = (0,\ 2,\ 1)^t$.

d.2) De $G^t\ x = y$, isto é, de:

$$\begin{pmatrix} 2 & 1 & -2 \\ & 3 & 2 \\ \bigcirc & & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix} ,$$

segue que:

$$x_3 = 1,$$

$$3\,x_2 + 2\,x_3 = 2 \Rightarrow x_2 = 0,$$

$$2\,x_1 + x_2 - 2\,x_3 = 0 \Rightarrow x_1 = 1.$$

Logo a solução do sistema $G^t x = y$ é $x(1,\ 0,\ 1)^t$.

Assim, a solução de $Ax = b$, isto é, de:

$$\begin{pmatrix} 4 & 2 & -4 \\ 2 & 10 & 4 \\ -4 & 4 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 6 \\ 5 \end{pmatrix} \text{ é } x = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} .$$

Exercícios

4.15 - *Aplicando-se o processo de Cholesky à matriz A, obteve-se:*

$$A = \begin{pmatrix} \dots & 2 & \dots & \dots \\ \dots & 8 & 10 & -8 \\ 3 & 10 & 14 & -5 \\ \dots & -8 & \dots & 29 \end{pmatrix} = GG^t ,$$

onde:

$$G = \begin{pmatrix} 1 & & & \bigcirc \\ 2 & \dots & & \\ \dots & 2 & 1 & \\ 0 & -4 & \dots & 2 \end{pmatrix} .$$

Preecher os espaços pontilhados com valores adequados.

4.16 - *Considere as matrizes:*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & -1 & 3 \end{pmatrix}; \quad B = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 2 \\ 0 & 2 & 1 \end{pmatrix}.$$

Escolha adequadamente e resolva um dos sistemas $Ax = b$, $Bx = b$, pelo processo de Cholesky, onde $b = (2,\ 1,\ 5)^t$.

4.17 - *Mostre que: Se o sistema de equações algébricas $Ax = b$, onde A é matriz não singular, é transformado no sistema equivalente $Bx = c$, com $B = A^tA$; $c = A^tb$, onde A^t é a transposta de A, então o último sistema pode sempre ser resolvido pelo processo de Cholesky (isto é, a matriz B satisfaz as condições para a aplicação do método). Aplicar a técnica acima para determinar pelo processo de Cholesky, a solução do sistema:*

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}.$$

4.6 Método de Eliminação de Gauss com Pivotamento Parcial

Além dos problemas já citados nesse capítulo para os métodos baseados na decomposição LU, existe um outro problema mais sério que está relacionado com a *propagação dos erros de truncamento* do computador. Assim, para ilustrar tal situação, consideremos um exemplo hipotético, (sistema linear de ordem 2, com uma máquina que trabalha apenas com 3 dígitos significativos). Tal exemplo servirá para ilustrar o que acontece com um sistema de grande porte num computador qualquer, visto que os mesmos operam com um número fixo e finito de algarismos significativos.

Exemplo 4.7 - *Através do método de Eliminação de Gauss, resolver o sistema linear:*

$$\begin{cases} 0.0001\, x_1 & + & 1.00\, x_2 & = & 1.00 \\ 1.00\, x_1 & + & 1.00\, x_2 & = & 2.00 \end{cases}$$

usando em todas as operações três dígitos significativos.

Solução: É fácil verificar que a solução desse sistema é:

$$x = \begin{pmatrix} 1.00010 \\ 0.99990 \end{pmatrix}.$$

Agora, resolvendo o sistema dado, pelo método de Eliminação de Gauss, com **3 dígitos significativos em todas as operações**, obtemos:

$$\left(\begin{array}{cc|c} 0.000100 & 1.00 & 1.00 \\ 1.00 & 1.00 & 2.00 \end{array} \right) \simeq \left(\begin{array}{cc|c} 0.000100 & 1.00 & 1.00 \\ & -10000 & -10000 \end{array} \right),$$

cuja solução é:

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

Portanto, obtemos uma solução muito diferente da solução exata do sistema.

A propagação de erros ocorre principalmente quando multiplicamos um número muito grande por outro que já contém erro de arredondamento. Por exemplo, suponha que um dado número z tenha um

erro de arredondamento ε; este número pode então ser escrito na forma: $\tilde{z} = z+\varepsilon$. Se agora multiplicamos esse número por p, então obtemos $p\tilde{z} = pz + p\varepsilon$ e assim o erro no resultado será $p\varepsilon$. Assim, se p for um número grande esse erro poderá ser muito maior que o original. Dizemos nesse caso que o erro em z foi amplificado.

No método de Eliminação de Gauss vários produtos com os multiplicadores são efetuados. Análises de propagação de erros de arredondamento para o algoritmo de Gauss indicam a conveniência de serem todos os multiplicadores (as constantes $a_{ik}^{(k)}/a_{kk}^{(k)}$ do k^{o} passo) menores que 1 em módulo; ou seja o pivô deve ser o elemento de maior valor absoluto da coluna, da diagonal (inclusive) para baixo.

Podemos então em cada passo, escolher na coluna correspondente o elemento de maior valor absoluto, da diagonal (inclusive) para baixo, e fazer uma permutação nas equações do sistema, de modo que esse elemento venha a ocupar a posição de pivô. A este procedimento chamamos **Método de Eliminação de Gauss com Pivotamento Parcial**.

Exemplo 4.8 - *Resolver o sistema do exemplo anterior pelo método de Eliminação de Gauss com pivotamento parcial.*

Solução: Devemos colocar na posição do pivô o elemento de maior valor absoluto da primeira coluna. Fazendo isso e aplicando o método de Eliminação de Gauss, obtemos:

$$\left(\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ 0.000100 & 1.00 & 1.00 \end{array} \right) \simeq \left(\begin{array}{cc|c} 1.00 & 1.00 & 2.00 \\ & 1.00 & 1.00 \end{array} \right) ,$$

cuja solução é:

$$x = \begin{pmatrix} 1.00 \\ 1.00 \end{pmatrix} .$$

e portanto bem mais próxima da solução exata.

A **matriz de Hilbert** é famosa por produzir um exemplo de sistema linear que se não utilizarmos pivotamento a solução obtida pode estar completamente errada. Os elementos de tal matriz são dados por:

$$h_{ij} = \frac{1}{i+j-1} , \quad i = 1, 2, \ldots, n ; \quad j = 1, 2, \ldots, n . \tag{4.11}$$

Assim a matriz de Hilbert de ordem 4 é dada por:

$$\begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix} .$$

Exercício

4.18 - *Considere um sistema linear de ordem 12 que tem a matriz de Hilbert como matriz dos coeficientes e que a solução exata seja o vetor que possui todas as componentes iguais a 1. Resolva o sistema usando:*

a) *o método de Eliminação de Gauss.*

b) *o método de Eliminação de Gauss com pivotamento parcial.*

Quem resolveu o exercício 4.18 (e quem não resolveu deve fazê-lo), pode observar que a solução obtida em b) é melhor que em a). No entanto, se no exercício 4.18 considerarmos um sistema de ordem 17, veremos que mesmo o método com pivotamento não produz uma solução satisfatória. O problema é que a matriz de Hilbert é uma matriz muito mal condicionada. O problema de condicionamento de matrizes será visto mais adiante.

4.7 Refinamento da Solução

Como já dissemos anteriormente, os métodos exatos deveriam fornecer, com um número finito de operações, a solução exata do sistema linear. Entretanto, devido aos erros de arredondamento obtemos, em geral, soluções aproximadas. Veremos aqui como refinar uma solução obtida por processo numérico.

Consideremos o seguinte sistema linear de ordem n:

$$\sum_{j=1}^{n} a_{ij}\ x_j \ = \ b_i\ ; \quad i = 1, 2, \ldots, n. \tag{4.12}$$

Sejam $x_1, x_2, \ldots, x_n$, a solução exata de (4.12) e sejam $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n$, uma aproximação da solução, por exemplo, obtida pelo método de Eliminação de Gauss. Observe que estamos considerando este método para poder descrever o processo de refinamento, mas o mesmo é válido para os demais métodos apresentados neste capítulo.

Então, devemos ter:

$$x_j \ = \ \bar{x}_j + y_j\ ; \quad j = 1, 2, \ldots, n, \tag{4.13}$$

onde y_j, j = 1, 2, ..., n, são as correções que devem ser adicionadas aos valores $\bar{x}_j$ (obtidos pelo método de Eliminação de Gauss), para fornecerem os valores corretos x_j. Substituindo (4.13) em (4.12) obtemos:

$$\begin{aligned} \sum_{j=1}^{n} a_{ij}\ (\bar{x}_j + y_j) \ &= \ b_i\ ; \quad i = 1, 2, \ldots, n, \\ \Rightarrow \sum_{j=1}^{n} a_{ij}\ y_j \ &= \ b_i - \sum_{j=1}^{n} a_{ij}\ \bar{x}_j\ ; \quad i = 1, 2, \ldots, n. \end{aligned}$$

Seja

$$r_i \ = \ b_i - \sum_{j=1}^{n} a_{ij}\ \bar{x}_j; \quad i = 1, 2, \ldots, n, \tag{4.14}$$

os resíduos. Obtemos então que:

$$\sum_{j=1}^{n} a_{ij}\ y_j \ = \ r_i; \quad i = 1, 2, \ldots, n. \tag{4.15}$$

Observações:

i) De (4.15), notamos que as correções são obtidas resolvendo-se um sistema linear análogo ao (4.12), com mesma matriz dos coeficientes e com o termo independente b_i, $i = 1, 2, \ldots, n$, substituído pelos resíduos r_i, $i = 1, 2, \ldots, n$. Assim na solução de (4.15), devemos refazer apenas os cálculos referentes ao novo termo independente. Portanto devemos, ao resolver (4.12), armazenar os multiplicadores (ou seja as constantes $a_{ik}^{(k)}/a_{kk}^{(k)}$), nas posições supostamente zeradas.

ii) A solução de (4.15) pode também estar afetada de erro, visto que as correções nada mais são do que a solução de sistema linear. Assim, encontrado y_j, substituimos seus valores em (4.13) e encontramos uma melhor solução aproximada $\bar{\bar{x}}$ à qual poderemos novamente aplicar o processo de refinamento. Obtemos com isso um processo iterativo, o qual deve ser aplicado até que uma precisão ϵ, pré-fixada, seja atingida.

iii) Para sabermos se a precisão foi atingida devemos calcular o vetor resíduo. Se o mesmo for o vetor nulo então teremos encontrado a solução exata. Se o vetor resíduo for diferente do vetor nulo então paramos o processo quando:

$$\frac{\| r^{(k+1)} - r^{(k)} \|_\infty}{\| r^{(k+1)} \|_\infty} < \epsilon \qquad (4.16)$$

onde $\| \, . \, \|_\infty$ está definida no Capítulo 1.

iv) No cálculo dos resíduos, ou seja, no cálculo de (4.14) haverá perda de algarismos significativos, por serem os valores de b_i e $\sum_{j=1}^{n} a_{ij} \, \bar{x}_j$ aproximadamente iguais. Assim devemos calcular os resíduos com precisão maior do que a utilizada nos demais cálculos. Logo, se estivermos trabalhando em *ponto flutuante* devemos calcular os resíduos em precisão dupla.

v) O processo de refinamento é utilizado, em geral, em sistemas de grande porte, onde o acúmulo de erros de arredondamento é maior.

O exemplo a seguir, mostra o processo de refinamento num caso hipotético, (sistema linear de ordem 2, com uma máquina que trabalha apenas com dois dígitos significativos). Novamente tal exemplo servirá para ilustrar o que acontece com um sistema de grande porte num computador qualquer.

Exemplo 4.9 - *Considere o sistema linear:*

$$\begin{pmatrix} 16. & 5.0 \\ 3.0 & 2.5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 21. \\ 5.5 \end{pmatrix}$$

Trabalhando com arredondamento para dois dígitos significativos em todas as operações;

a) *resolva o sistema pelo método de Eliminação de Gauss,*

b) *faça uma iteração para refinar a solução obtida em* **a)**.

Solução: Aplicando o método de Eliminação de Gauss ao sistema dado, obtemos:

$$\left(\begin{array}{cc|c} 16. & 5.0 & 21. \\ 3.0 & 2.5 & 5.5 \end{array}\right) \sim \left(\begin{array}{c|c|c} 16. & 5.0 & 21. \\ \hline 0.19 & 1.6 & 1.5 \end{array}\right),$$

Observe que não calculamos o elemento $a_{21}^{(2)}$, mas armazenamos nesta posição o multiplicador $\frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{3.0}{16.} = 0.1875 = 0.19$, o qual deverá ser usado no cálculo do novo termo independente. Assim, resolvendo o sistema:

$$\begin{pmatrix} 16. & 5.0 \\ 0 & 1.6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 21. \\ 1.5 \end{pmatrix},$$

obtemos:

$$\begin{aligned} x_2 &= \frac{1.5}{1.6} = 0.9375 = 0.94 \\ x_1 &= \frac{21. - 5.0 \times 0.94}{16.} = \frac{21. - 4.7}{16.} = \frac{16.3}{16.} = \frac{16.}{16.} = 1.0 \end{aligned}$$

Portanto, a solução aproximada é: $\bar{x} = (1.0, 0.94)^t$. Agora, desde que $r = b - A\bar{x}$, obtemos:

$$r = \begin{pmatrix} 21. \\ 5.5 \end{pmatrix} - \begin{pmatrix} 16. & 5.0 \\ 3.0 & 1.6 \end{pmatrix} \begin{pmatrix} 1.0 \\ 0.94 \end{pmatrix}.$$

Fazendo os cálculos, (lembrando que devemos utilizar precisão dupla), obtemos:

$$r = \begin{pmatrix} 21.00 - 16.00 - 4.700 \\ 5.500 - 3.000 - 2.350 \end{pmatrix} = \begin{pmatrix} 0.3000 \\ 0.1500 \end{pmatrix}.$$

Assim, consideramos :

$$r = \begin{pmatrix} 0.30 \\ 0.15 \end{pmatrix}$$

Devemos apenas refazer os cálculos em relação ao novo termo independente, isto é:

$$r = \begin{pmatrix} 0.30 \\ 0.15 \end{pmatrix} \sim \begin{pmatrix} 0.30 \\ 0.15 - 0.30 \times 0.19 \end{pmatrix}$$

$$\Rightarrow r = \begin{pmatrix} 0.30 \\ 0.15 - 0.057 \end{pmatrix} = \begin{pmatrix} 030 \\ 0.093 \end{pmatrix}$$

Assim, resolvendo o sistema:

$$\begin{pmatrix} 16. & 5.0 \\ 0 & 1.6 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0.30 \\ 0.093 \end{pmatrix}$$

obtemos: $y = (0.00063,\ 0.058)^t$. Fazendo,

$$x = \bar{x} + y$$

segue que:

$$x = \begin{pmatrix} 1.0 \\ 0.94 \end{pmatrix} + \begin{pmatrix} 0.00063 \\ 0.058 \end{pmatrix} = \begin{pmatrix} 1.00063 \\ 0.998 \end{pmatrix} = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}.$$

Calculando novamente o resíduo obtemos que o mesmo é o vetor nulo. Assim $x = (1.0, 1.0)^t$ é a solução exata do sistema dado.

Exercício

4.19 -*Considere o sistema linear:*

$$\begin{cases} x_1 + 3x_2 + 4x_3 = -5 \\ 3x_1 + 2x_2 + x_3 = 8 \\ 2x_1 + 4x_2 + 3x_3 = 4 \end{cases}$$

a) *Resolva-o pelo método de Eliminação de Gauss, trabalhando com arredondamento para dois dígitos significativos em todas as operações.*

b) *Refine a solução obtida em* **a)**.

4.20 - *Considere o sistema linear:*

$$\begin{cases} 2x_1 + 3x_2 + 4x_3 = -2 \\ 3x_1 + 2x_2 - x_3 = 4 \\ 5x_1 - 4x_2 + 3x_3 = 8 \end{cases}$$

a) *Resolva-o pelo método de Eliminação de Gauss com pivotamento parcial, trabalhando com arredondamento para dois dígitos significativos em todas as operações.*

b) *Refine a solução obtida em* **a)**.

4.8 Mal Condicionamento

Como vimos, para resolver sistemas lineares dois aspectos devem ser considerados:

a) se a solução existe ou não.

b) achar um modo eficiente para resolver as equações.

Mas existe ainda um aspecto a ser considerado:

c) se a solução das equações é muito sensível a pequenas mudanças nos coeficientes.

Este fenômeno é chamado **Mal Condicionamento**, e está relacionado ao fato de que a matriz dos coeficientes nas equações lineares está *próxima* de ser singular.

Na seção anterior dissemos que se o vetor resíduo for próximo do vetor nulo então a solução obtida está razoavelmente precisa, e isto é verdade para sistemas *bem condicionados*. Entretanto em alguns casos, como mostrado no exemplo a seguir, isto está longe de ser verdadeiro.

Exemplo 4.10 - *Considere o sistema linear:*

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix},$$

e suponha dada a solução aproximada:

$$\bar{x} = \begin{pmatrix} 0.9911 \\ -0.4870 \end{pmatrix}.$$

Calcule o vetor resíduo.

Solução: Calculando o vetor resíduo correspondente a $\bar{x}$, através de (4.14), obtemos:

$$r = \begin{pmatrix} 10^{-8} \\ -10^{-8} \end{pmatrix}$$

e portanto parece razoável supor que o erro em $\bar{x}$ é muito pequeno. Entretanto, pode ser verificado por substituição que a solução exata é $x = (2,\ 2)^t$. No caso deste exemplo é fácil reconhecer o extremo mal condicionamento do sistema. De fato, o elemento

$$a_{22}^{(2)} = 0.1441 - 0.8648 \times \frac{0.2161}{1.2969} = 0.1441 - 0.1440999923 \simeq 10^{-8}.$$

Assim uma pequena mudança no elemento 0.1441 resultará numa grande mudança em $a_{22}^{(2)}$ e portanto em x_2, ou seja, mudando 0.1441 para 0.1442 teremos $a_{22}^{(2)} = 0.0001000077$ com solução

$$\bar{x} = \begin{pmatrix} -0.00015399 \\ 0.66646092 \end{pmatrix}.$$

Portanto, a menos que os coeficientes em A e b sejam dados com uma precisão melhor do que 10^{-8} é perigoso falar sobre uma solução do sistema dado.

Observe que com este exemplo, não queremos dizer que todas as soluções aproximadas de equações mal condicionadas fornecem resíduos pequenos, mas apenas que algumas soluções aproximadas de equações mal condicionadas fornecem resíduos bem pequenos.

Para a análise da pertubação, é conveniente sermos capazes de associar a qualquer vetor ou matriz um escalar não negativo que em algum sentido mede suas grandezas. Tais medidas que satisfazem alguns axiomas são chamadas normas. (Revise normas de vetores e normas de matrizes, Capítulo 1).

Análise da Pertubação

Vamos investigar, agora, a *condição* de um sistema linear não singular $Ax = b$. Desde que A é não singular, a solução do sistema é dada por: $x = A^{-1}b$.Vamos supor aqui que os dados estão sujeitos a certas pertubações e vamos analisar o efeito dessas pertubações na solução. Seja x a solução exata do sistema $Ax = b$.

1º **caso:** Consideremos uma pertubação do vetor b da forma $b + \delta b$, e seja A conhecida exatamente. Portanto a solução x também será pertubada, isto é, teremos $x + \delta x$, e assim de:

$$A(x + \delta x) = b + \delta b \; . \tag{4.17}$$

obtemos:

$$(x + \delta x) \; = \; A^{-1}(b + \delta b) \; . \tag{4.18}$$

A questão que queremos resolver é como relacionar δx com δb, ou seja, sabendo o tamanho da pertubação em δb como estimar a pertubação em δx? O procedimento a seguir responde essa pergunta. De (4.17), obtemos:

$$Ax + A\delta x \; = \; b + \delta b \; \Rightarrow \; A\delta x \; = \; \delta b \; ,$$

desde que $Ax = b$. Agora desde que A é não singular, segue que:

$$\delta x \; = \; A^{-1}\delta b \; .$$

Aplicando norma, em ambos os membros, e usando normas consistentes, obtemos:

$$\| \delta x \| \leq \| A^{-1} \| \| \delta b \| \; . \tag{4.19}$$

Do mesmo modo, de $Ax = b$, obtemos:

$$\| b \| \leq \| A \| \| x \| \; . \tag{4.20}$$

Multiplicando, membro a membro, (4.19) por (4.20), obtemos:

$$\| \delta x \| \| b \| \leq \| A \| \| A^{-1} \| \| \delta b \| \| x \| \; , \tag{4.21}$$

ou

$$\frac{\| \delta x \|}{\| x \|} \leq \| A \| \| A^{-1} \| \frac{\| \delta b \|}{\| b \|} \; .$$

Assim a pertubação relativa em x está relacionada com a pertubação relativa em b pela constante multiplicativa $\| A \| \| A^{-1} \|$.

Definindo, o **número de condição de** A, como:

$$cond(A) \; = \| A \| \| A^{-1} \| \; ,$$

obtemos:

$$\frac{\| \delta x \|}{\| x \|} \leq \; cond(A) \; \frac{\| \delta b \|}{\| b \|} \; .$$

Observações:

a) Temos que $cond(A) \geq 1$. De fato:

$$cond(A) \; = \| A \| \| A^{-1} \| \geq \| A \; A^{-1} \| = \| I \| = 1 \; .$$

b) $\dfrac{\| \delta b \|}{\| b \|}$ pode ser interpretada como uma medida do erro relativo em b. O erro em $\dfrac{\| \delta x \|}{\| x \|}$ dependerá do valor do número de condição que é maior ou igual a 1.

c) Se $cond(A)$ é grande, então pequenas pertubações relativas em b produzirão grandes pertubações relativas em x, e o problema de resolver $Ax = b$ é mal condicionado.

d) $cond(A)$ será considerado grande quando valer por volta de 10^4 ou mais.

2º **caso:** Consideremos agora uma pertubação da matriz A da forma $A + \delta A$ e seja b conhecido exatamente. Portanto a solução x também será também pertubada, isto é, teremos $x + \delta x$, e assim de:

$$(A + \delta A)(x + \delta x) \; = \; b \; , \tag{4.22}$$

obtemos:

$$(x + \delta x) \; = \; (A + \delta A)^{-1} b \; . \tag{4.23}$$

Mas, $x = A^{-1}b$. Portanto:

$$\delta x \; = \; -A^{-1}b + (A + \delta A)^{-1}b \; \Rightarrow \; \delta x \; = \; [(A + \delta A)^{-1} - A^{-1}]b \; .$$

Seja $B = A + \delta A$. Temos que:

$$B^{-1} - A^{-1} \; = \; A^{-1}AB^{-1} - A^{-1}BB^{-1} \; = \; A^{-1}(A - B)B^{-1} \; .$$

Logo:

$$\begin{aligned} \delta x \; &= \; [A^{-1}(A - B)B^{-1}]b \; = \; [A^{-1}(A - (A + \delta A))(A + \delta A)^{-1}]b \\ \Rightarrow \; \delta x \; &= \; -A^{-1}\delta A(A + \delta A)^{-1}b \; . \end{aligned}$$

De (4.23), segue que:

$$\delta x \; = \; -A^{-1}\delta A(x + \delta x) \; .$$

Aplicando norma, em ambos os membros, e usando normas consistentes, obtemos:

$$\begin{aligned} \| \delta x \| \; &\leq \; \| A^{-1} \| \| \delta A \| \| x + \delta x \| \\ \Rightarrow \frac{\| \delta x \|}{\| x + \delta x \|} \; &\leq \; \| A^{-1} \| \| A \| \frac{\| \delta A \|}{\| A \|} \\ \Rightarrow \frac{\| \delta x \|}{\| x + \delta x \|} \; &\leq \; cond(A) \frac{\| \delta A \|}{\| A \|} \; . \end{aligned}$$

Novamente, se $cond(A)$ é grande, então pequenas pertubações em A, produzirão grandes pertubações relativas em x e o problema de resolver $Ax = b$ é mal condicionado.

Exemplo 4.11 - *Analisar o sistema linear:*

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.8642 \\ 0.1440 \end{pmatrix} .$$

Solução: Temos que :

$$A^{-1} \ = \ 10^8 \begin{pmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{pmatrix} .$$

(O cálculo da matriz inversa se encontra na próxima seção).

Usando norma linha, obtemos:

$$\| A \|_\infty = \ 2.1617 \ , \quad \| A^{-1} \|_\infty = \ 1.5130 \times 10^8 \ ,$$

e portanto

$$cond(A) \ = \| A \|_\infty \| A^{-1} \|_\infty = \ 327065210 \ \simeq 3.3 \times 10^8 \ ,$$

mostrando que o sistema é extremamente mal condicionado.

Se for de interesse estudar métodos que sejam particularmente úteis no caso da matriz A ser mal condicionada, citamos aqui o **método de Kaczmarz**, que pode ser encontrado por exemplo em [Carnahan, 1969].

Exercícios

4.21 - *Prove que se uma norma de matrizes é subordinada a uma norma do $\mathbb{R}^n$ elas são consistentes.*

4.22 - *Sejam A e B matrizes de ordem n. Prove que:*

$$\frac{\| B^{-1} - A^{-1} \|}{\| B^{-1} \|} \leq cond(A) \frac{\| A - B \|}{\| B \|} .$$

4.23 - *Analisar o sistema linear $Ax = b$, onde:*

$$A \ = \ \begin{pmatrix} 100 & 99 \\ 99 & 98 \end{pmatrix} .$$

4.24 - *Analisar o sistema linear $Ax = b$, de ordem 17, onde os elementos de A são dados por (4.11), isto é, A é a matriz de Hilbert.*

4.9 Cálculo da Matriz Inversa

Sejam A uma matriz não singular ($det(A) \neq 0$), $A^{-1} = [b_1 \vdots b_2 \vdots \ldots \vdots b_n]$ a matriz inversa de A, onde b_j é a coluna j da matriz A^{-1} e e_j a coluna j da matriz identidade. De $AA^{-1} = I$; isto é, de :

$$A \left[b_1 \vdots b_2 \vdots \ldots \vdots b_n \right] \ = \ \left[e_1 \vdots e_2 \vdots \ldots \vdots e_n \right],$$

resulta:

$$Ab_j \ = \ e_j; \quad j = 1, 2, \ldots, n.$$

Assim podemos calcular as colunas j, $j = 1, 2, \ldots, n$, da matriz A^{-1}, resolvendo os sistemas lineares acima.

Portanto, podemos inverter uma matriz utilizando qualquer um dos métodos dados nesse Capítulo.

Observações:

1) Usando decomposição LU obtemos as colunas de A^{-1}, fazendo:

$$LUb_i \ = \ e_i \ , \qquad i = 1, 2, \ldots, n \ ,$$

isto é, resolvendo os sistemas:

$$\begin{cases} L \ y_i \ = \ e_i \\ U \ b_i \ = \ y_i \end{cases} \quad i = 1, 2, \ldots, n \ .$$

2) Usando o método de Cholesky (somente para matrizes simétricas e positivas definidas), obtemos as colunas de A^{-1}, fazendo:

$$GG^t b_i \ = \ e_i \ , \qquad i = 1, 2, \ldots, n \ ,$$

isto é, resolvendo os sistemas:

$$\begin{cases} G \ y_i \ = \ e_i \\ G^t \ b_i \ = \ y_i \end{cases} \quad i = 1, 2, \ldots, n \ .$$

3) Usando o método de Eliminação de Gauss, obtemos as colunas de A^{-1}, resolvendo os sistemas:

$$Ab_i \ = \ e_i \quad , \quad i = 1, 2, \ldots, n.$$

Observe que podemos colocar todas as colunas da identidade ao lado da matriz A e fazer a decomposição de uma só vez.

4) Podemos calcular a inversa de uma matriz, pelo método de Gauss-Compacto usando o mesmo esquema da resolução de sistemas matriciais, isto é, fazendo:

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \ldots & & & \\ x_{n1} & x_{n2} & \ldots & x_{nn} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & & & \\ 0 & 0 & \ldots & 1 \end{pmatrix} .$$

Portanto, as colunas da matriz X são as colunas da matriz inversa de A, desde que $AA^{-1} = I$.

Exemplo 4.12 - *Considere a matriz:*

$$A \ = \ \begin{pmatrix} 3 & 0 & 3 \\ 2 & -2 & 1 \\ 1 & 2 & 0 \end{pmatrix} .$$

Calcule A^{-1} utilizando o Método de Gauss-Compacto.

Solução: Devemos resolver o sistema matricial:

$$\begin{pmatrix} 3 & 0 & 3 \\ 2 & -2 & 1 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Temos:

$$\left(\begin{array}{ccc|ccc} 3 & 0 & 3 & 1 & 0 & 0 \\ 2 & -2 & 1 & 0 & 1 & 0 \\ 1 & 2 & 0 & 0 & 0 & 1 \end{array}\right) \sim \left(\begin{array}{ccc|ccc} 3 & 0 & 3 & 1 & 0 & 0 \\ 2/3 & -2 & -1 & -2/3 & 1 & 0 \\ 1/3 & -1 & -2 & -1 & 1 & 1 \end{array}\right) .$$

Resolvendo o sistema:

$$\begin{pmatrix} 3 & 0 & 3 \\ & -2 & -1 \\ \bigcirc & & -2 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix} = \begin{pmatrix} 1 \\ -2/3 \\ -1 \end{pmatrix} ,$$

obtemos: $x_{11} = -\frac{1}{6}$, $x_{21} = \frac{1}{12}$, $x_{31} = \frac{1}{2}$, (que é a 1ª coluna de A^{-1}). Do mesmo modo:

$$\begin{pmatrix} 3 & 0 & 3 \\ & -2 & -1 \\ \bigcirc & & -2 \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} ,$$

fornece $x_{12} = \frac{1}{2}$, $x_{22} = -\frac{1}{4}$, $x_{32} = -\frac{1}{2}$, (que é a 2ª coluna de A^{-1}), e de:

$$\begin{pmatrix} 3 & 0 & 3 \\ & -2 & -1 \\ \bigcirc & & -2 \end{pmatrix} \begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} ,$$

segue que: $x_{13} = \frac{1}{2}$, $x_{23} = \frac{1}{4}$, $x_{33} = -\frac{1}{2}$, (que é a 3ª coluna de A^{-1}). Assim:

$$A^{-1} = \begin{pmatrix} -1/6 & 1/2 & 1/2 \\ 1/12 & -1/4 & 1/4 \\ 1/2 & -1/2 & -1/2 \end{pmatrix} .$$

Exercícios

4.25 - *Usando decomposição LU, inverter a matriz:*

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} .$$

4.26 - *Dada a matriz:*

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 10 & 2 \\ -1 & 2 & 4 \end{pmatrix} ,$$

calcular A^{-1}, utilizando o processo de Cholesky.

4.27 - *Seja*

$$A = \begin{pmatrix} 2 & 4 & 6 \\ 1 & -3 & -1 \\ 2 & 1 & 1 \end{pmatrix} .$$

Usando o método de Eliminação de Gauss, calcule A^{-1}.

4.28 - *Usando o método de Gauss-Compacto, calcule A^{-1}, onde:*

$$A = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 0 & 2 \\ 4 & -1 & 3 \end{pmatrix} .$$

4.10 Exercícios Complementares

4.29 - *Quais das matrizes:*

$$A = \begin{pmatrix} 2 & 2 & 1 \\ 3 & 3 & 2 \\ 3 & 2 & 1 \end{pmatrix}; \quad B = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 2 & 1 \\ 3 & 3 & 2 \end{pmatrix}; \quad C = \begin{pmatrix} 2 & 1 & 3 \\ 4 & 3 & 8 \\ 6 & 7 & 17 \end{pmatrix},$$

podem ser decompostas na forma LU ? Decompor as que forem possíveis.

4.30 - *Mostre que se A é uma matriz real, simétrica, positiva definida, então necessariamente temos:*

a) $a_{ii} > 0, \; i = 1, 2, \ldots, n$.

b) $a_{ik}^2 < a_{ii} \, a_{kk}$, *para todo* $i \neq k$.

c) *o maior elemento de A em módulo está sob a diagonal.*

4.31 - *Considere o sistema* $Ax = b$*; onde*

$$A = \begin{pmatrix} 1 & \alpha & 3 \\ \alpha & 1 & 4 \\ 5 & 2 & 1 \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}; \quad b = \begin{pmatrix} -2 \\ -3 \\ 4 \end{pmatrix}.$$

Para que valores de α*:*

i) *A matriz A é decomponível no produto LU ? Justifique.*

ii) *O sistema pode ser resolvido por Cholesky . Justifique.*

iii) *Considere* $\alpha = 1$*, e resolva o sistema pelo método de Eliminação de Gauss .*

4.32 - *Resolva o sistema abaixo pelo método de Cholesky, completando adequadamente os espaços pontilhados.*

$$\begin{cases} 2\,x_1 + \ldots\,x_2 - x_3 = 3 \\ x_1 + 10\,x_2 + \ldots\,x_3 = 6 \\ \ldots\,x_1 + 2\,x_2 + 4\,x_3 = -6 \end{cases}$$

4.33 - *Para que valores de* α *e* β *a matriz:*

$$\begin{pmatrix} 4 & \alpha & 1 \\ \beta & 4 & 1 \\ 1 & 1 & 1 \end{pmatrix},$$

se decompõe no produto GG^t.

4.34 - *Considere os sistemas:*

$$(I) \begin{cases} x_1 + 2\,x_2 - x_3 = 4 \\ 2\,x_1 + 13\,x_2 + x_3 = 35 \\ -\,x_1 + x_2 + 4\,x_3 = 5 \end{cases}$$

$$(II) \begin{cases} x_1 + 2\,x_2 + x_3 = 6 \\ 2\,x_1 + 4\,x_2 + x_3 = 14 \\ 2\,x_1 + 2\,x_2 + x_3 = 6 \end{cases}$$

Faça uma escolha adequada para resolver um deles pelo método de Gauss-Compacto e o outro pelo método de Cholesky. Justifique sua resposta.

4.35 - *Resolva o seguinte sistema, por Eliminação de Gauss, usando aritmética complexa.*

$$\begin{cases} (2+3i)\, x + (2-i)\, y = 2+i \\ (4+6i)\, x + (3-6i)\, y = -2-5i \end{cases}$$

4.36 - *No exercício anterior escreva $x = x_r + ix_i$; $y = y_r + iy_i$. Multiplique as partes real e imaginária de cada equação separadamente. Mostre que o resultado é um sistema de 4 equações a 4 incógnitas, cuja solução são as partes real e imaginária do exercício anterior.*

4.37 - *Se a decomposição LU de uma matriz simétrica A é dada por:*

$$L = \begin{pmatrix} 1 & & & \bigcirc \\ 2 & 1 & & \\ 3 & 2 & 1 & \\ 4 & 3 & 2 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1 & 2 & 3 & 4 \\ & 4 & 8 & 12 \\ \bigcirc & & 9 & 18 \\ & & & 16 \end{pmatrix},$$

verifique que a matriz G da decomposição Cholesky é dada por:

$$G = \begin{pmatrix} 1 & & & \bigcirc \\ 2 & 2 & & \\ 3 & 4 & 3 & \\ 4 & 6 & 6 & 4 \end{pmatrix}.$$

4.38 - *Resolver o sistema matricial:*

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 & | & y_1 \\ x_2 & | & y_2 \\ x_3 & | & y_3 \end{pmatrix} = \begin{pmatrix} 4 & | & 2 \\ 2 & | & -2 \\ 9 & | & 7 \end{pmatrix}.$$

pelo método de Gauss-Compacto.

4.39 - *Calcular u_2, u_3, u_4, u_5 resolvendo a equação de diferenças:*

$$u_{n+2} + 4u_{n+1} + u_n = n\ , \quad (*)$$

com as condições de contorno $u_1 = 0$ e $u_6 = 1$, usando um método a sua escolha.(Escreva $(*)$ para $n = 1, 2, 3, 4$).

4.40 - *Aplicando-se o método de Cholesky a uma matriz foi obtido:*

$$A = \begin{pmatrix} 1 & 2 & -1 \\ \dots & 13 & \dots \\ \dots & 1 & 4 \end{pmatrix} = GG^t \quad \text{onde} \quad G = \begin{pmatrix} \dots & 0 & \dots \\ 2 & \dots & \dots \\ -1 & \dots & \sqrt{2} \end{pmatrix}.$$

a) *Preencha os espaços pontilhados com valores adequados.*

b) *Usando a decomposição GG^t, calcule a inversa de A.*

4.41 - *Seja o sistema $Ax = b$, dado por:*

$$\begin{pmatrix} 10 & 7 & 8 \\ 7 & 5 & 6 \\ 8 & 6 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -3 \\ -1 \\ 7 \end{pmatrix}.$$

a) *Determine a inversa da matriz dos coeficientes pelo método de Eliminação de Gauss.*

b) *Resolva o sistema dado, utilizando a matriz inversa obtida em* **a)**.

c) *Determine a solução do sistema usando o método de Gauss-Compacto.*

4.42 *Relacione os sistemas lineares:*

$$(I)\left\{\begin{array}{rcrcrcr} & & 3\,x_2 & + & 2\,x_3 & = & 5 \\ x_1 & + & 4\,x_2 & + & x_3 & = & 6 \\ & & 2\,x_2 & + & 5\,x_3 & = & 7 \end{array}\right.$$

$$(II)\left\{\begin{array}{rcrcrcr} -2\,x_1 & + & 2\,x_2 & & & = & -1 \\ x_1 & + & 3\,x_2 & - & x_3 & = & 3 \\ & - & x_2 & + & 2\,x_3 & = & 1 \end{array}\right.$$

$$(III)\left\{\begin{array}{rcrcrcr} x_1 & + & 2\,x_2 & + & x_3 & = & 4 \\ 2\,x_1 & + & 6\,x_2 & & & = & 8 \\ x_1 & & & + & 4\,x_3 & = & 5 \end{array}\right.$$

com os métodos:

A) *Eliminação de Gauss,*

B) *Cholesky,*

C) *Gauss-Compacto,*

para a sua resolução.

4.43 - *Considere o seguinte conjunto esparso de equações:*

$$\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & \bigcirc & \\ & -1 & 2 & -1 & & \\ & & -1 & 2 & -1 & \\ & \bigcirc & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 7 \\ 5 \\ 4 \\ 3 \end{pmatrix}$$

Mostre que usando o método de Eliminação de Gauss o sistema triangular resultante permanece esparso. Um sistema como este é chamado **TRIDIAGONAL**. *Tais sistemas aparecem frequentemente na solução de equações diferenciais parciais.*

4.44 - *Considere o sistema:*

$$\begin{pmatrix} 2 & 5 & 3 \\ 5 & 2 & 1 \\ 1 & 3 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 8 \\ 7 \\ 13 \end{pmatrix} .$$

Trabalhando com arredondamento para dois dígitos significativos em todas as operações:

a) *Resolva o sistema acima pelo método de Eliminação de Gauss com pivotamento parcial.*

b) *Faça uma iteração para refinar a solução obtida em* **a)** *e então calcule o vetor resíduo. O que você pode concluir?*

4.45 - *Dado o sistema $Ax = b$, considere uma pertubação da matriz A da forma $A + \delta A$ e seja b conhecido exatamente. Prove que: se*

$$\| A^{-1} \| \leq \frac{1}{1-\delta} \| B^{-1} \| ,$$

com $B = A + \delta A$, então:

$$\| \delta x \| \leq \frac{\delta}{1-\delta} \| x + \delta x \| \ , \delta = \| \delta A \| \| B^{-1} \| < 1 \ .$$

4.46 - *Considere um sistema linear cuja matriz dos coeficientes é dada por:*

$$A = \begin{pmatrix} \epsilon & -1 & 1 \\ -1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} ,$$

onde $\epsilon << 1$.

a) *Calcule o número de condição de A.*

b) *Com base no resultado do item* **a)** *a aplicação do método de Eliminação de Gauss daria bom resultado ou seria necessário usar Eliminação de Gauss com pivotamento parcial?*

c) *Considere $\epsilon = 10^{-4}$, e resolva o sistema linear $Ax = b$, onde $b = (2,\ 0,\ 1)^t$, pelo método escolhido no item* **b)**.

4.47 - *No Capítulo 3, vimos que o processo iterativo:*

$$x_{k+1} = x_k \ (2 - a \ x_k) \ , \ a \neq 0 \ ,$$

pode ser usado para obter $\frac{1}{a}$.

A fórmula acima vale também para refinar a inversa A^{-1} de uma matriz A não singular, isto é: dada uma aproximação inicial X_0 de A^{-1}, podemos refinar esta aproximação usando:

$$X_{k+1} = X_k \ (2 \ I - A \ x_k) \ , \ k = 0, 1, \ldots \qquad (**)$$

Considere a matriz:

$$\begin{pmatrix} 3.00 & 1.00 \\ 2.00 & 2.00 \end{pmatrix} .$$

Trabalhando com arredondamento para três dígitos significativos em todas as operações;

a) *obtenha uma aproximação para A^{-1}, usando o método de Eliminação de Gauss,*

b) *refine a inversa obtida em* **a)** *usando (**), até obter o resíduo $R_k = I - A \ X_k$ com norma < 0.1,*

c) *usando a inversa obtida em* **b)** *calcule a solução aproximada do sistema $Ax = b$, onde $b = (-14.0,\ 12.0)^t$.*

4.48 - *Seja A uma matriz não singular de ordem n, e sejam u e v vetores n-dimensionais.*

a) *Mostre que, se $(A - uv^t)^{-1}$ existe, então:*

$$(A - uv^t)^{-1} = A^{-1} + \alpha A^{-1} u v^t A^{-1} \quad \text{com} \alpha = \frac{1}{1 - v^t A^{-1} u} \ .$$

b) *Dê condições para a existência da inversa* $(A - uv^t)^{-1}$.

c) *Se* A^{-1} *é conhecida, e* B *é uma matriz que coincide com* A*, exceto em uma linha, podemos escolher* u *e* v *para obter* B^{-1} *(se existir), aplicando a fórmula dada no item* **a)**. *Sabendo que:*

$$A = \begin{pmatrix} 12 & -4 & 7 \\ -4 & 1 & -2 \\ 7 & -2 & 4 \end{pmatrix} \;, \quad A^{-1} = \begin{pmatrix} 0 & -2 & -1 \\ -2 & 1 & 4 \\ -1 & 4 & 4 \end{pmatrix} \;,$$

e que B *coincide com* A *exceto que em vez de 12 temos 5, calcule* B^{-1}.

4.11 Problemas Aplicados e Projetos

4.1 - *Considere o circuito a seguir com resistências e baterias tal como indicado; escolhemos arbitrariamente as correntes e os valores da malha:*

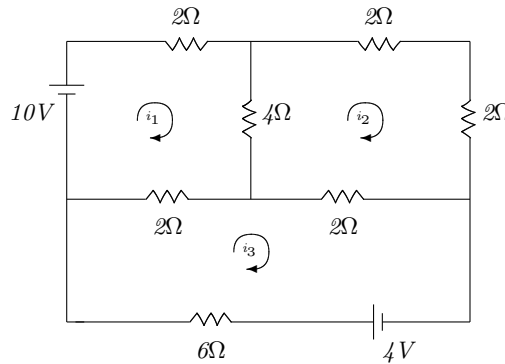



Figura 4.4

Aplicando a Lei de Kirchoff que diz que a soma algébrica da diferenças de potencial em qualquer circuito fechado é zero, obtemos para as correntes i_1, i_2, i_3*, o seguinte sistema linear:*

$$\begin{cases} 2\,i_1 \;+\; 4\,(i_1 - i_2) \;+\; 2\,(i_1 - i_3) \;-\; 10 & = \; 0 \\ 2\,i_2 \;-\; 2\,i_2 \;+\; 2\,(i_2 - i_3) \;+\; 4\,(i_2 - i_1) & = \; 0 \\ 6\,i_3 \;+\; 2\,(i_3 - i_1) \;+\; 2\,(i_3 - i_2) \;-\; 4 & = \; 0 \end{cases}$$

Deseja-se determinar o valor de $i = (i_1,\ i_2,\ i_3)^t$ *que satisfaça o sistema acima.*

a) *É possível resolver o sistema pelo método da decomposição LU? Justifique.*

b) *É possível resolver o sistema pelo método de Cholesky? Justifique.*

c) *Resolva o sistema pelo método de Eliminação de Gauss.*

4.2 - *Representemos por* x_1, x_2, x_3 *e* x_4 *o número de quatro produtos que podem ser produzidos no decorrer de uma semana. Para a produção de cada unidade precisa-se de três tipos diferentes de matéria prima A, B e C conforme indicado na Tabela 4.1.*

Tabela 4.1

matéria prima / *Produto*	A	B	C
(1)	1	2	4
(2)	2	0	1
(3)	4	2	3
(4)	3	1	2

Por exemplo: para produzir uma unidade de (1) precisa-se de 1 unidade de A, 2 de B e 4 de C. Se existem disponíveis 30,20 e 40 unidades respectivamente de A, B e C, quantas unidades de cada produto podemos produzir?

Obs: Escreva x_1, x_2 *e* x_3 *em função de* x_4 *e lembre-se que as soluções devem ser inteiras e não negativas.*

4.3 - *Um cachorro está perdido em um labirinto quadrado de corredores (Figura 4.5). Em cada interseção escolhe uma direção ao acaso e segue até a interseção seguinte onde escolhe novamente ao acaso nova direção e assim por diante. Qual a probabilidade do cachorro, estando na interseção i, sair eventualmente pelo lado sul?*

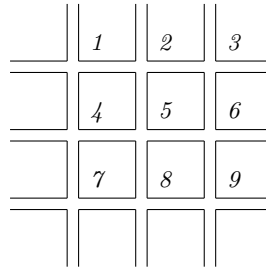



Figura 4.5

Esclarecimentos: *Suponhamos que há exatamente as 9 interseções mostradas na figura. Seja* P_1 *a probabilidade do cachorro, que está na interseção 1, sair pelo lado sul. Seja* $P_2, P_3, \ldots, P_9$ *definidas de modo similar. Supondo que em cada interseção a que chegue o cachorro, há tanta possibilidade que escolha uma direção como outra, e que, tendo chegado a uma saída, tenha terminado sua caminhada, a teoria das probabilidades oferece as seguintes equações para* P_i*:*

$$\left\{ \begin{array}{lcl} P_1 & = & (0 + 0 + P_2 + P_4)/4 \\ P_2 & = & (0 + P_1 + P_3 + P_5)/4 \\ P_3 & = & (0 + P_2 + 0 + P_6)/4 \\ P_4 & = & (P_1 + 0 + P_5 + P_7)/4 \\ P_5 & = & (P_2 + P_4 + P_6 + P_8)/4 \\ P_6 & = & (P_3 + P_5 + 0 + P_9)/4 \\ P_7 & = & (P_4 + 0 + P_8 + 1)/4 \\ P_8 & = & (P_5 + P_7 + P_9 + 1)/4 \\ P_9 & = & (P_6 + P_8 + 0 + 1)/4 \end{array} \right.$$

Para saber a resposta resolva o sistema linear obtido.

4.4 - *O problema de se determinar um polinômio:*

$$P_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \ ,$$

de grau no máximo n, tal que :

$$\int_a^b x^i P_n(x)dx = k_i, \quad i = 0, 1, \ldots, n \ ,$$

onde k_i são constantes, pode ser resolvido através da obtenção da solução de um sistema linear. Determine um polinômio de grau 3, que satisfaça a condição acima, considerando $a = -1$, $b = 1$ e:

a) $k_0 = \frac{2}{3}, \quad k_1 = \frac{4}{3}, \quad k_2 = \frac{6}{5}, \quad k_3 = \frac{4}{5}$,

b) $k_0 = 2, \quad k_1 = 2, \quad k_2 = \frac{2}{3}, \quad k_3 = \frac{58}{35}$,

resolvendo o sistema linear resultante por método numérico à sua escolha.

4.5 - *Cargas horizontal e vertical X e Y e um momento M são aplicados à uma estrutura em balanço de comprimento L, como mostrado na Figura 4.6.*

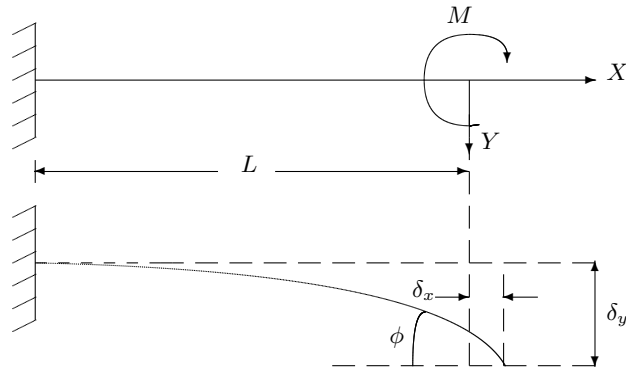



Figura 4.6

Na extremidade livre, o alongamento δx, a deflexão δy e o giro ϕ (igual ao ângulo com a horizontal) são relacionados com as cargas, da seguinte maneira:

$$\begin{pmatrix} \frac{L}{E\ A} & 0 & 0 \\ 0 & \frac{L^3}{3\ E\ I} & \frac{L^2}{2\ E\ I} \\ 0 & \frac{L^2}{2\ E\ I} & \frac{L}{E\ I} \end{pmatrix} \begin{pmatrix} X \\ Y \\ M \end{pmatrix} = \begin{pmatrix} \delta x \\ \delta y \\ \phi \end{pmatrix} ,$$

onde:

- *E é o módulo de Young*
- *A é a àrea de secção transversal*
- *I é o momento de inércia*

A matriz do sistema acima é positiva definida e é chamada matriz de flexibilidade da estrutura.

a) *Obter a inversa da matriz de flexibilidade correspondente aos seguintes dados:*

- $E = 200 \frac{t}{cm^2}$
- $A = 400\ cm^2$
- $L = 3.0\ m$
- $I = 50000\ cm^4$

b) *Usando o resultado obtido em* **a)** *calcular as cargas X e Y e o momento M correspondentes a:*

- $\delta x = 0.0035\ cm$
- $\delta y = 3.0\ cm$
- $\phi = 0.018$

4.6 - *Considere o circuito em escada, mostrado na Figura 4.7, composto de resistores concentrados e fontes de tensão ideais, uma independente e outra controlada.*

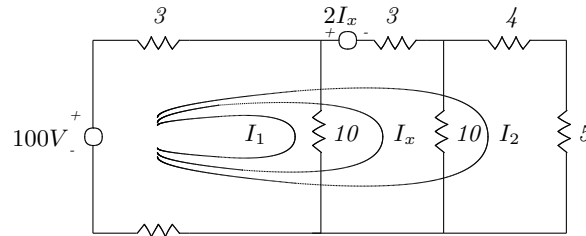



Figura 4.7

O sistema de equações que se obtém ao solucionar o circuito pelo método dos laços é o seguinte:

$$\begin{cases} 4\ (I_1 + I_2 + I_x) + 10\ I_1 = 100 \\ 4\ (I_1 + I_2 + I_x) + 3\ (I_2 + I_x) + 10\ I_x = 100 - 2\ I_x \\ 4\ (I_1 + I_2 + I_x) + 3\ (I_2 + I_x) + 9\ I_x = 100 - 2\ I_x \end{cases}$$

a) *Escreva o sistema acima na forma $Ax = b$.*

b) *Determine A^{-1}.*

c) *Usando o resultado obtido em* **b)** *determine o valor de I_x.*

4.7 - *Suponha que tenhamos o circuito dado na Figura 4.8.*

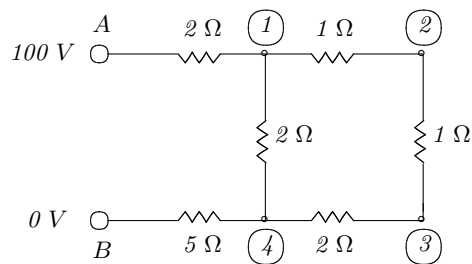



Figura 4.8

A corrente que flui do nó p para o nó q de uma rede elétrica é dada por:

$$I_{pq} = \frac{V_p - V_q}{R_{pq}}$$

onde I em Ampéres, R em Ohms e V_p e V_q são voltagens nos nós p e q, respectivamente, e R_{pq} é a resistência no arco pq (Lei de Ohm).

A soma das correntes que chegam a cada nó é nula (Lei de Kirchoff); assim, as equações que relacionam as voltagens podem ser obtidas. Por exemplo: no nó 1, tem-se a equação :

$$I_{A1} + I_{21} + I_{41} = 0,$$

ou seja,

$$\frac{100 - V_1}{2} + \frac{V_2 - V_1}{1} + \frac{V_4 - V_1}{2} = 0$$

ou ainda

$$-4V_1 + 2V_2 + V_4 = -100.$$

Obtenha as demais equações do sistema linear e resolva-o.

4.8 - *Uma transportadora possui 5 tipos de caminhões que representaremos por (1), (2), (3), (4), (5), os quais são equipados para transportar 5 tipos diferentes de máquinas A, B, C, D, E segundo a Tabela 4.2, onde supomos que A, B, C, D, E é a quantidade de máquinas que cada caminhão pode transportar levando carga plena.*

Tabela 4.2

Máquinas / *Caminhões*	*A*	*B*	*C*	*D*	*E*
(1)	*1*	*1*	*1*	*0*	*2*
(2)	*0*	*1*	*2*	*1*	*1*
(3)	*2*	*1*	*1*	*2*	*0*
(4)	*3*	*2*	*1*	*2*	*1*
(5)	*2*	*1*	*2*	*3*	*1*

Assim, o caminhão (1) pode transportar 1 máquina A, 1 máquina B, 1 máquina C, nenhuma máquina D, 2 máquina E, etc. Quantos caminhões de cada tipo devemos enviar para transportar exatamente:

- *27 máquinas do tipo A*
- *23 máquinas do tipo B*
- *31 máquinas do tipo C*
- *31 máquinas do tipo D*
- *22 máquinas do tipo E*

Supondo que cada caminhão vai com carga plena, resolva o sistema linear obtido.

Sugestão: Represente por x_1, x_2, x_3, x_4 e x_5 o número de caminhões respectivamente dos tipos (1), (2), (3), (4) e (5).

4.9 - *Elabore um algoritmo que tendo como dados* n*, a matriz* $A(n \times n)$ *onde* $A = (a_{ij})$ *é tal que* $a_{ij} = 0$ *para* $|i - j| > 1$*, e* b *é um vetor* $(n \times 1)$*, determina a solução do sistema linear* $Ax = b$ *pelo método de Eliminação de Gauss adaptado para sistemas tridiagonais.*

a) *Teste seu algoritmo para resolver o sistema dado por:*

$$-y_{k-1} + 2y_k - y_{k+1} = \frac{8}{(n+1)^2}, \quad k = 1, 2, ..., n,$$

para vários valores de n*, e compare sua solução com a solução matemática:*

$$y_k = 4\left[\frac{k}{n+1} - \left(\frac{k}{n+1}\right)^2\right].$$

Considere $n = 10$ *e* $n = 20$ *e tome em ambos os casos* $y_0 = y_{n+1} = 0$.

b) *Teste seu algoritmo para resolver o sistema* $Ax = b$*, onde:*

$$A = \begin{pmatrix} 4 & -2 & & & & & & & & \\ -1 & 2 & -1 & & & & \bigcirc & & & \\ & -1 & 2 & -1 & & & & & & \\ & & -1 & 2 & -1 & & & & & \\ & & & -1 & 2 & -1 & & & & \\ & & & & -1 & 2 & -1 & & & \\ & & & & & -1 & 2 & -1 & & \\ & & & & & & -1 & 2 & -1 & \\ & \bigcirc & & & & & & -1 & 2 & -1 \\ & & & & & & & & -2 & 4 \end{pmatrix}, \quad e\,,$$

$b = (2,\ -1,\ 7,\ 5,\ 4,\ 3,\ 2,\ -4,\ 7,\ 8)^t$.

4.10 - *Representemos por* x_i*,* $i = 1, 2, \ldots, n$ *o número das unidades de* n *produtos que podem ser produzidos no decorrer de uma semana. Para produção de cada unidade precisa-se de* m *tipos diferentes de matéria prima* $M_1, M_2, \ldots, M_m$*. Tais relações são dadas através de uma matriz* A *onde* a_{ij} *indica quantas unidades da matéria prima* M_j *são necessárias para produzir uma unidade do produto* x_i*.*

Suponha que existam $D_1, D_2, \ldots, D_m$ *unidades respectivamente de matérias primas* $M_1, M_2, \ldots, M_m$*. Nosso problema é determinar quantas unidades de cada produto podemos produzir.*

Lembre-se que tais quantidades devem ser inteiras e não negativas.

Considere os dados da Tabela 4.3.

Tabela 4.1

	M_1	M_2	M_3	M_4	M_5	M_6
x_1	1	2	4	2	4	5
x_2	2	0	1	2	2	2
x_3	4	2	3	1	5	3
x_4	3	1	2	3	2	2
x_5	1	2	0	3	1	2
x_6	1	0	1	0	4	3
x_7	5	3	2	2	3	2
D	60	40	40	50	70	60

Capítulo 5

Solução de Sistemas Lineares: Métodos Iterativos

5.1 Introdução

Ao lado dos métodos exatos para resolver sistemas lineares, existem os métodos iterativos os quais passamos a discutir agora. Em certos casos, tais métodos são melhores do que os exatos, por exemplo, quando a matriz dos coeficientes é uma matriz esparsa (muitos elementos iguais a zero). Eles ainda são mais econômicos no sentido que utilizam menos memória do computador. Além disso, possuem a vantagem de se auto corrigir se um erro é cometido, e eles podem ser usados para reduzir os erros de arredondamento na solução obtida por métodos exatos, como discutido no Capítulo 4. Podem também sob certas condições ser aplicado para resolver um conjunto de equações não lineares.

5.2 Processos Estacionários.

Um método é **iterativo** quando fornece uma sequência de aproximantes da solução, cada uma das quais obtida das anteriores pela repetição do mesmo tipo de processo.

Um método iterativo é **estacionário** se cada aproximante é obtido do anterior sempre pelo mesmo processo.

Quando os processos variam de passo para passo mas se repetem ciclicamente de s em s passos dizemos que o processo é s**-cíclico**. Agrupando-se os s passos de cada ciclo num único passo composto, obtemos um método estacionário.

No caso de métodos iterativos precisamos sempre saber se a sequência que estamos obtendo está convergindo ou não para a solução desejada. Além disso, precisamos sempre ter em mente o significado de convergência. (Revise: norma de vetor e norma de matriz , Capítulo 1).

Definição 5.1 - *Dados uma sequência de vetores $x^{(k)} \in E$ e uma norma sobre E, onde E é um espaço vetorial, dizemos que a sequência $\{x^{(k)}\}$ converge para $x \in E$ se $\| x^{(k)} - x \| \rightarrow 0$, quando $k \rightarrow \infty$.*

Consideremos então um sistema linear $Ax = b$ determinado, ($det(A) \neq 0$), onde A é uma matriz quadrada de ordem n, x e b são vetores $n \times 1$.

Como no caso de equações não lineares (Capítulo 3), para determinar a solução de um sistema linear por métodos iterativos, precisamos transformar o sistema dado em um outro sistema onde possa ser definido um processo iterativo; e mais, que a solução obtida para o sistema transformado seja também solução do sistema original, isto é, os sistemas lineares devem ser equivalentes.

Suponha então, que o sistema $Ax = b$ tenha sido transformado num sistema equivalente da forma:

$$x \; = \; B\,x \; + \; g \; . \tag{5.1}$$

(por exemplo: $B = I - A$ e $g = b$), de maneira que a solução $\bar{x}$ de (5.1) seja, também, solução de $Ax = b$.

Seja $x^{(0)}$ uma aproximação inicial para a solução $\bar{x}$ de (5.1). Obtemos as aproximações sucessivas $x^{(k)}$ para a solução desejada $\bar{x}$, usando o processo iterativo estacionário definido por:

$$x^{(k)} \; = \; B\,x^{(k-1)} \; + \; g \; . \tag{5.2}$$

Se a sequência $\{x^{(k)}\}$ converge para $\bar{x}$ então $\bar{x}$ coincide com a solução x de $Ax = b$.

De fato, passando-se ao limite ambos os membros de (5.2), obtêm-se:

$$\bar{x} = B\;\bar{x} \; + \; g \; .$$

Pela hipótese de equivalência $\bar{x}$ é também solução de $Ax = b$.

O próximo teorema fornece a condição necessária e suficiente para a convergência da sequência $x^{(k)}$.

Teorema 5.1 - *A condição* **necessária e suficiente** *para a convergência do processo iterativo definido por (5.2) é que* $max|\lambda_i| < 1$*, onde* λ_i *são os auto-valores da matriz* B*.*

Prova: A prova deste teorema pode ser encontrada em .....

Em geral é difícil verificar as condições do teorema 5.1. Entretanto, podemos obter uma condição **suficiente** que a matriz B deve satisfazer para assegurar a convergência do processo iterativo definido por (5.2). Enunciamos formalmente tal condição no próximo corolário.

Corolário 5.1 - **(Critério Geral de convergência)**- *O processo iterativo definido por (5.2) é convergente se para qualquer norma de matrizes,* $\parallel B \parallel < 1$*.*

Prova: A convergência da sequência $x^{(k)}$ para a solução x de $Ax = b$ é estudada introduzindo-se o vetor erro:

$$e^{(k)} = x - x^{(k)} \; .$$

Subtraindo-se (5.2) membro a membro de (5.1), obtemos:

$$x \; - \; x^{(k)} \; = \; B \; \left(x - x^{(k-1)} \right) \; ,$$

e portanto:

$$e^{(k)} \; = \; B\; e^{(k-1)} \; . \tag{5.3}$$

De (5.3) podemos escrever:

$$e^{(k-1)} \; = \; B\; e^{(k-2)} \; \Rightarrow \; e^{(k)} \; = \; B^2\; e^{(k-2)} \; ,$$

e assim por aplicações sucessivas, segue que:

$$e^{(k)} \; = \; B^k\; e^{(0)},$$

onde $e^{(0)}$ é o erro inicial. Tomando normas consistentes, (definição 1.13), na expressão acima, segue que:

$$\parallel B^k\; e^{(0)} \parallel \leq \parallel B \parallel^k \parallel e^{(0)} \parallel \; .$$

Portanto:

$$\| e^k \| \leq \| B \|^k \| e^{(0)} \| \ .$$

Desta desigualdade vemos que se $\| B \| < 1$ teremos:

$$\| e^{(k)} \| = \| x - x^{(k)} \| \to 0 \ ,$$

isto é, se $\| B \| < 1$ para alguma norma, então temos garantida a convergência do processo iterativo definido por (5.2).

A matriz B de (5.2) é chamada **matriz de iteração** do processo iterativo.

Exemplo 5.1 - *Seja*

$$B = \begin{pmatrix} 0.5 & -0.2 & 0.5 \\ 0.1 & 0.6 & 0.4 \\ -0.3 & 0.1 & 0.0 \end{pmatrix} .$$

Verificar se um sistema $Ax = b$ que tenha a matriz B acima, como matriz de iteração, convergirá para a solução.

Solução: Calculando $\| B \|_\infty$ obtemos $\| B \|_\infty = 1.2$ e nada podemos concluir. Calculando $\| B \|_1$ obtemos $\| B \|_1 = 0.9 < 1$ e podemos agora afirmar que o processo iterativo com essa matriz é convergente.

Processo de Parada

Para aplicar qualquer método iterativo escolhemos $x^{(0)}$ como uma aproximação inicial para a solução de $Ax = b$. Com essa aproximação inicial e um método numérico, do tipo (5.2), *refinamos* a solução até obtê-la com uma determinada precisão (número de casas decimais corretas) .

Para obtermos a solução com uma determinada precisão ϵ devemos, durante o processo iterativo, efetuar o seguinte teste: Se

$$\frac{\| x^{(k+1)} - x^{(k)} \|_\infty}{\| x^{(k+1} \|_\infty} < \epsilon \ (errorelativo) \ ,$$

onde ϵ é uma precisão pré-fixada; x^k e x^{k+1} são duas aproximações consecutivas para $\bar{x}$, então x^{k+1} é a solução procurada, isto é, tomamos $x = x^{k+1}$.

Veremos agora alguns métodos particulares.

5.2.1 Método de Jacobi-Richardson

Considere o sistema linear $Ax = b$ de ordem n, determinado, isto é:

$$\left\{ \begin{array}{lllllllll} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \ldots & + & a_{2n}x_n & = & b_2 \\ \ldots \ldots & & & & & & & & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \ldots & + & a_{nn}x_n & = & b_n \end{array} \right. \tag{5.4}$$

A matriz A do sistema (5.4) pode ser decomposta na forma:

$$A = L + D + R \ ,$$

onde L é uma matriz triangular inferior formada pela parte inferior da matriz A, D é a diagonal de A e R é uma matriz triangular superior formada pela parte superior da matriz A, isto é:

$$\ell_{ij} = \left\{ \begin{array}{l} a_{ij} \ , \ i > j \\ 0 \ , \ i \leq j \end{array} \right. \ ; \ d_{ij} = \left\{ \begin{array}{l} a_{ij} \ , \ i = j \\ 0 \ , \ i \neq j \end{array} \right. \ ; \ r_{ij} = \left\{ \begin{array}{ll} a_{ij} & i < j \\ 0 & i \geq j \end{array} \right. .$$

Supondo $det(D) \neq 0$, podemos transformar o sistema original em:

$$\begin{aligned} &(L + D + R)\, x = b \\ &\Rightarrow Dx = -\,(L+R)\,x + b \\ &\Rightarrow x = -D^{-1}\,(L+R)\,x + D^{-1}\,b\,. \end{aligned}$$

que está na forma (5.1) com $B = -D^{-1}(L+R)$ e $g = D^{-1}b$.

O processo iterativo definido por:

$$x^{(k+1)} = -D^{-1}(L+R)\,x^{(k)} + D^{-1}\,b, \tag{5.5}$$

é chamado de **Método de Jacobi-Richardson**.

Comparando (5.5) com (5.2) vemos que a matriz de iteração do método de Jacobi-Richardson é : $-D^{-1}(L+R)$.

Por hipótese $a_{ii} \neq 0$, pois estamos supondo $det(D) \neq 0$. Podemos então, antes de decompor a matriz A em $L+D+R$, dividir cada equação pelo correspondente elemento da diagonal principal, resultando assim:

$$A^* = L^* + I + R^*\,,$$

onde A^* é a matriz obtida de A após a divisão, I é a matriz identidade.

Assim, o processo iterativo pode ser escrito como:

$$x^{(k+1)} = -\,(L^* + R^*)\;x^{(k)} + b^*, \tag{5.6}$$

onde os elementos de L^*, R^* e b^* são, respectivamente, dados por:

$$\ell^*_{ij} = \begin{cases} a^*_{ij} = \dfrac{a_{ij}}{a_{ii}}, & i > j \\ 0\,, & i \leq j \end{cases} \;;\quad r^*_{ij} = \begin{cases} a^*_{ij} = \dfrac{a_{ij}}{a_{ii}}, & i < j \\ 0\,, & i \geq j \end{cases} \;;$$

$$b^*_i = \frac{b_i}{a_{ii}}\,,\; i = 1, 2, \ldots, n\,.$$

Vemos por (5.6) que as componentes de x^{k+1} podem ser calculadas sucessivamente sem necessidade de se calcular D^{-1}, e a matriz de iteração do método de Jacobi-Richardson é dada por: $-\,(L^* + R^*)$.

Dado o sistema (5.4), o método de Jacobi-Richardson consiste na determinação de uma sequência de aproximantes da iteração k:

$$x_1^{(k)}\,,\; x_2^{(k)}\,,\; \ldots\,,\; x_n^{(k)} \quad,\quad k = 1, 2, 3, \ldots\,,$$

a partir de valores iniciais:

$$x_1^{(0)}\,,\; x_2^{(0)}\,,\; \ldots\,,\; x_n^{(0)}\,,$$

através do processo iterativo definido por (5.6), isto é:

$$\begin{cases} x_1^{(k+1)} = -\,a^*_{12}x_2^{(k)} - a^*_{13}x_3^{(k)} - \ldots - a^*_{1n}x_n^{(k)} + b^*_1 \\ x_2^{(k+1)} = -\,a^*_{21}x_1^{(k)} - a^*_{23}x_3^{(k)} - \ldots - a^*_{2n}x_n^{(k)} + b^*_2 \\ x_3^{(k+1)} = -\,a^*_{31}x_1^{(k)} - a^*_{32}x_2^{(k)} - \ldots - a^*_{3n}x_n^{(k)} + b^*_3 \\ \ldots\;\ldots \\ x_n^{(k+1)} = -\,a^*_{n1}x_1^{(k)} - a^*_{n2}x_2^{(k)} - \ldots - a^*_{n,n-1}x_{n-1}^{(k)} + b^*_n \end{cases}$$

Observe que o método de iteração linear para uma única equação, que foi discutido no Capítulo 3, e o método de Jacobi-Richardson são exatamente o mesmo, com a diferença que este último é aplicado

a um sistema de equações. É fácil ver que no método de Jacobi-Richardson as equações são mudadas simultaneamente usando-se o valor mais recente do vetor x, e por causa disso esse método é também conhecido por **Método dos Deslocamentos Simultâneos**.

Critérios de Convergência

Fazendo $B = -(L^* + R^*)$ no critério geral de convergência, (Lema 5.1) e escolhendo sucessivamente as normas $\| \cdot \|_\infty$ e $\| \cdot \|_1$ obtemos critérios suficientes de convergência para o método de Jacobi-Richardson. Assim o método de Jacobi-Richardson **converge** se:

a) o **critério das linhas** for satisfeito, isto é, se:

$$\max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}^*| \; < \; 1 \; , \tag{5.7}$$

b) o **critério das colunas** for satisfeito, isto é, se:

$$\max_{1 \le j \le n} \sum_{\substack{i=1 \\ i \ne j}}^{n} |a_{ij}^*| \; < \; 1 \; , \tag{5.8}$$

Observe que basta apenas um dos critérios ser satisfeito para garantirmos a convergência.

Definição 5.2 - *Uma matriz A é* **estritamente diagonalmente dominante** *se:*

$$\sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}| \; < \; |a_{ii}| \; , \quad i = 1, 2, \ldots, n \; . \tag{5.9}$$

Observação:

Com a definição 5.2 fica fácil ver que se a matriz dos coeficientes for estritamente diagonalmente dominante então o critério das linhas é satisfeito. De fato, se (5.9) é satisfeita para todo i, então se dividimos cada equação pelo correspondente elemento da diagonal principal segue que:

$$\sum_{\substack{j=1 \\ j \ne i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1, i = 1, \ldots, n \Rightarrow \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}^*| < 1, i = 1, \ldots, n \Rightarrow \max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \ne i}}^{n} |a_{ij}^*| < 1 \; .$$

Portanto obtemos mais um critério de convergência, isto é, podemos verificar se o método de jacobi-Richardson converge testando se a matriz dos coeficientes é estritamente diagonalmente dominante.

Exemplo 5.2 - *Resolver o sistema:*

$$\left\{ \begin{array}{rcrcrcr} 10x_1 & + & 2x_2 & + & x_3 & = & 7 \\ x_1 & + & 5x_2 & + & x_3 & = & -8 \\ 2x_1 & + & 3x_2 & + & 10x_3 & = & 6 \end{array} \right. .$$

pelo método de Jacobi-Richardson com $x^{(0)} = (0.7, -1.6, 0.6)^t$ *e* $\epsilon < 10^{-2}$.

Solução: Em primeiro lugar devemos testar se temos garantia de convergência. Temos que a matriz dos coeficientes é estritamente diagonalmente dominante, pois satisfaz (5.9). De fato:

$$\begin{array}{rcl} |a_{12}| + |a_{13}| & = & |2| + |1| \ < \ |10| \ = \ |a_{11}| \ , \\ |a_{21}| + |a_{23}| & = & |1| + |1| \ < \ |5| \ = \ |a_{22}| \ , \\ |a_{31}| + |a_{32}| & = & |2| + |3| \ < \ |10| \ = \ |a_{33}| \ . \end{array}$$

Portanto podemos garantir que o processo de Jacobi-Richardson aplicado ao sistema dado será convergente.

Dividindo então cada equação pelo correspondente elemento da diagonal principal obtemos:

$$\left\{ \begin{array}{ccccccc} x_1 & + & 0.2x_2 & + & 0.1x_3 & = & 0.7 \\ 0.2x_1 & + & x_2 & + & 0.2x_3 & = & -1.6 \\ 0.2x_1 & + & 0.3x_2 & + & x_3 & = & 0.6 \end{array} \right.$$

Apesar de não ser necessário, pois já sabemos que o processo de Jacobi-Richardson será convergente, por se tratar de exemplo, verificaremos também o critério das linhas e o critério das colunas. Assim, para verificar o critério das linhas, calculamos:

$$\begin{array}{l} |a^*_{12}| + |a^*_{13}| \ = \ |0.2| + |0.1| \ = \ 0.3 \ , \\ |a^*_{21}| + |a^*_{23}| \ = \ |0.1| + |0.1| \ = \ 0.2 \ , \\ |a^*_{31}| + |a^*_{32}| = \ |0.2| + |0.3| \ = \ 0.5 \ , \\ \Rightarrow \max\limits_{1 \leq i \leq n} \sum\limits_{\substack{j=1 \\ j \neq i}}^{3} |a^*_{ij}| \ = \ 0.5 \ < \ 1 \ , \end{array}$$

e portanto o critério das linhas é válido, o que já era de se esperar, (ver observação após definição 5.2).

Para verificar o critério das colunas, calculamos:

$$\begin{array}{l} |a^*_{21}| + |a^*_{31}| \ = \ |0.2| + |0.2| \ = \ 0.4 \ , \\ |a^*_{12}| + |a^*_{32}| \ = \ |0.2| + |0.3| \ = \ 0.5 \ , \\ |a^*_{13}| + |a^*_{23}| \ = \ |0.1| + |0.2| \ = \ 0.3 \ , \\ \Rightarrow \ \max\limits_{1 \leq j \leq n} \sum\limits_{\substack{i=1 \\ i \neq j}}^{3} |a^*_{ij}| \ = \ 0.5 \ < \ 1 \ , \end{array}$$

portanto o critério das colunas também é válido.

Portanto qualquer um dos critérios de convergência assegura a convergência do método de Jacobi-Richardson. Temos que as iterações são definidas por:

$$\left\{ \begin{array}{ccccccc} x_1^{(k+1)} & = & -0.2x_2^{(k)} & - & 0.1x_3^{(k)} & + & 0.7 \\ x_2^{(k+1)} & = & -0.2x_1^{(k)} & - & 0.2x_3^{(k)} & - & 1.6 \\ x_3^{(k+1)} & = & -0.2x_1^{(k)} & - & 0.3x_2^{(k)} & + & 0.6 \end{array} \right. .$$

e a partir de $x^{(0)} = (0.7, -1.6, 0.6)^t$, obtemos para $x^{(1)}$ os seguintes valores:

$$\left\{ \begin{array}{ccl} x_1^{(1)} & = & -0.2x_2^{(0)} - 0.1x_3^{(0)} + 0.7 = -0.2(-1.6) - 0.1(0.6) + 0.7 = 0.96 \\ x_2^{(1)} & = & -0.2x_1^{(0)} - 0.2x_3^{(0)} - 1.6 = -0.2(0.7) - 0.2(0.6) - 1.6 = -1.86 \\ x_3^{(1)} & = & -0.2x_1^{(0)} - 0.3x_2^{(0)} + 0.6 = -0.2(0.6) - 0.3(-1.6) + 0.6 = 0.94 \end{array} \right. .$$

Continuando as iterações obtemos a tabela:

k	0	1	2	3	4
x_1	0.7	0.96	0.978	0.9994	0.9979
x_2	-1.6	-1.86	-1.98	-1.9888	-1.9996
x_3	0.6	0.94	0.966	0.9984	0.9968

Agora, desde que:

$$x^{(4)} - x^{(3)} = \begin{pmatrix} -0.0015 \\ 0.0108 \\ 0.0016 \end{pmatrix} ,$$

e portanto

$$\frac{\| x^{(4)} - x^{(3)} \|_\infty}{\| x^{(4)} \|_\infty} = \frac{0.0108}{1.9996} \simeq 0.0054 < 10^{-2} ,$$

segue que a solução do sistema, com $\epsilon < 10^{-2}$, é:

$$x = \begin{pmatrix} 0.9979 \\ -1.9996 \\ 0.9978 \end{pmatrix} .$$

Exercícios

5.1 - *Usando o método de Jacobi-Richardson, obter a solução do sistema:*

$$\begin{cases} 10x_1 + x_2 - x_3 = 10 \\ x_1 + 10x_2 + x_3 = 12 \\ 2x_1 - x_2 + 10x_3 = 11 \end{cases} .$$

com 3 casas decimais corretas.

5.2 - *Dado o sistema:*

$$\begin{cases} 10x_1 + x_2 - x_3 = 10 \\ 2x_1 + 10x_2 + 8x_3 = 20 \\ 7x_1 + x_2 + 10x_3 = 30 \end{cases} .$$

a) *Verificar a possiblidade de aplicação do método de Jacobi-Richardson.*

b) *Se possível resolvê-lo pelo método do item* **a)**, *obtendo o resultado com* $\epsilon < 10^{-2}$.

5.2.2 Método de Gauss-Seidel.

Suponhamos, como foi feito para o método de Jacobi-Richardson, que o sistema linear $Ax = b$ seja escrito na forma:

$$(L^* + I + R^*)x = b^* .$$

Transformamos então esse sistema como se segue:

$$\begin{aligned} &(L^* + I)x = -R^*x + b^* \\ &x = -(L^* + I)^{-1}R^*x + (L^* + I)^{-1}b^* . \end{aligned}$$

que está na forma (5.1) com $B = -(L^* + I)^{-1}R^*$ e $g = (L^* + I)^{-1}b^*$.

O processo iterativo definido por:

$$x^{(k+1)} = -(L^* + I)^{-1} R^* x^{(k)} + (L^* + I)^{-1} b^* \ , \tag{5.10}$$

é chamado de **Método de Gauss-Seidel**.

Comparando (5.10) com (5.2) vemos que a matriz de iteração do método de Gauss-Seidel é: $-(L^* + I)^{-1} R^*$.

Observe que pré-multiplicando (5.10) por $(L^* + I)$, segue que:

$$(L^* + I) x^{(k+1)} = -R^* x^{(k)} + b^*$$

ou

$$x^{(k+1)} = -L^* x^{(k+1)} - R^* x^{(k)} + b^* \ . \tag{5.11}$$

Por (5.11) vemos que as componentes de $x^{(k+1)}$ podem ser calculadas sucessivamente sem necessidade de se calcular $(L^* + I)^{-1}$.

Dado o sistema (5.4), o método de Gauss-Seidel consiste na determinação de uma sequência de aproximantes da iteração k:

$$x_1^{(k)} \ , \ x_2^{(k)} \ , \ldots , \ x_n^{(k)} \qquad k = 1, 2, 3, \ldots \ ,$$

a partir de valores iniciais:

$$x_1^{(0)} \ , \ x_2^{(0)} \ , \ldots , \ x_n^{(0)} \ ,$$

através do processo iterativo definido por (5.11), isto é:

$$\left\{ \begin{array}{lcl} x_1^{(k+1)} & = & - a_{12}^* x_2^{(k)} - a_{13}^* x_3^{(k)} - \ldots - a_{1n}^* x_n^{(k)} + b_1^* \\ x_2^{(k+1)} & = & - a_{21}^* x_1^{(k+1)} - a_{23}^* x_3^{(k)} - \ldots - a_{2n}^* x_n^{(k)} + b_2^* \\ x_3^{(k+1)} & = & - a_{31}^* x_1^{(k+1)} - a_{32}^* x_2^{(k+1)} - \ldots - a_{3n}^* x_n^{(k)} + b_3^* \\ \ldots \ldots & & \\ x_n^{(k+1)} & = & - a_{n1}^* x_1^{(k+1)} - a_{n2}^* x_2^{(k+1)} - \ldots - a_{n,n-1}^* x_{n-1}^{(k+1)} + b_n^* \end{array} \right.$$

Esse método difere do processo de Jacobi-Richardson por utilizar para o cálculo de uma componente de $x^{(k+1)}$ o valor mais recente das demais componentes. Por esse motivo o método da Gauss-Seidel também é conhecido por **Método dos Deslocamentos Sucessivos**.

Critérios de Convergência

Fazendo $B = -(L^* + I)^{-1} R^*$ no critério geral de convergência, (Lema 5.1), vamos agora obter critérios de convergência para o método de Gauss-Seidel.

Inicialmente lembremos, (veja definição 1.13), que se k satisfizer a desiguladade $\| Bx \| \leq k \ \ \| x \|$ teremos $\| B \| \leq k$. Impondo $k < 1$ teremos uma condição suficiente para garantir a convergência do método de Gauss-Seidel. Vejamos então como obter tal condição. Para tanto, seja $y = Bx$, isto é:

$$\begin{array}{ll} & y \ = \ -(L^* + I)^{-1} \ R^* x \\ \Rightarrow & (L^* + I)\, y \ = \ - R^* x \\ \Rightarrow & y = -L^* y \ - \ R^* x \ . \end{array}$$

Assim o vetor y é obtido do vetor x a partir das equações:

$$\left\{ \begin{array}{lcl} y_1 & = & - a_{12}^* \ x_2 - a_{13}^* \ x_3 - \ldots - a_{1n}^* \ x_n \\ y_2 & = & - a_{21}^* \ y_1 - a_{23}^* \ x_3 - \ldots - a_{2n}^* \ x_n \\ y_3 & = & - a_{31}^* \ y_1 - a_{32}^* \ y_2 - \ldots - a_{3n}^* \ x_n \\ \vdots & & \\ y_n & = & - a_{n1}^* \ y_1 - a_{n2}^* \ y_2 - \ldots - a_{n,n-1}^* \ y_{n-1} \end{array} \right. \tag{5.12}$$

e queremos calcular $\| Bx \|_\infty$. Mas,

$$\| Bx \|_\infty = \| y \|_\infty \ , \quad \text{com} \quad \| y \|_\infty = \max_{1 \leq i \leq n} |y_i| \ .$$

Agora, a partir de (5.12), podemos escrever:

$$\begin{aligned} |y_1| &= |\sum_{j=2}^{n} a^*_{1j} \ x_j \ | \ \leq \ \sum_{j=2}^{n} |a^*_{1j}| \ |x_j| \ \leq \ \sum_{j=2}^{n} |a^*_{1j}| \ \max_j \ |x_j| \\ &= \sum_{j=2}^{n} |a^*_{1j}| \ \| x \|_\infty = \beta_1 \ \| x \|_\infty \quad \text{onde} \quad \beta_1 \ = \ \sum_{j=2}^{n} |a^*_{1j}| \ . \end{aligned}$$

Portanto:

$$|y_1| \ \leq \ \beta_1 \ \| x \|_\infty \ .$$

$$\begin{aligned} |y_2| &= |a^*_{21} \ y_1 \ + \ \sum_{j=3}^{n} a^*_{2j} \ x_j| \ \leq \ |a^*_{21}||y_1| \ + \ \sum_{j=3}^{n} |a^*_{2j}| \ |x_j| \\ &\leq |a^*_{21}|\beta_1 \ \| x \|_\infty + \sum_{j=3}^{n} |a^*_{2j}| \ \max_j |x_j| \\ &\leq |a^*_{21}| \ \beta_1 \ \| x \|_\infty \ + \sum_{j=3}^{n} |a^*_{2j}| \ \| x \|_\infty \\ &= \left(|a^*_{21}|\beta_1 \ + \ \sum_{j=3}^{n} |a^*_{2j}| \right) \ ||x||_\infty \ = \ \beta_2 \ ||x||_\infty \\ & \quad \text{onde } \beta_2 \ = \ \left(|a^*_{21}|\beta_1 \ + \ \sum_{j=3}^{n} |a^*_{2j}| \right) \end{aligned}$$

Portanto:

$$|y_2| \ \leq \ \beta_2 \ ||x||_\infty \ .$$

Assim, para y_i, podemos escrever:

$$\begin{aligned} |y_i| &= |\sum_{j=1}^{i-1} a^*_{ij} \ y_j \ + \ \sum_{j=i+1}^{n} a^*_{ij} \ x_j| \ \leq \ \sum_{j=1}^{i-1} |a^*_{ij}| \ |y_j| \ + \ \sum_{j=i+1}^{n} |a^*_{ij}| \ |x_j| \\ &\leq \sum_{j=1}^{i-1} |a^*_{ij}| \ \beta_j \ \| x \|_\infty \ + \ \sum_{j=i+1}^{n} |a^*_{ij}| \max_j \ |x_j| \\ &\leq \left(\sum_{j=1}^{i-1} |a^*_{ij}| \ \beta_j \ + \ \sum_{j=j+1}^{n} |a^*_{ij}| \right) \ \| x \|_\infty = \ \beta_i \ \| x \|_\infty \\ & \quad \text{onde } \beta_i \ = \ \left(\sum_{j=1}^{i-1} |a^*_{ij}| \ \beta_j \ + \ \sum_{j=i+1}^{n} |a^*_{ij}| \right) \end{aligned}$$

Portanto:

$$|y_i| \leq \beta_i \, ||x||_\infty \, .$$

Temos então:

$$\begin{aligned} \| Bx \|_\infty &= \| y \|_\infty = \max_{1 \leq i \leq n} |y_i| \leq \max_{1 \leq i \leq n} \beta_i \; \| x \|_\infty \\ \Rightarrow & \quad \| B \|_\infty \leq \max_{1 \leq i \leq n} \beta_i \, . \end{aligned}$$

Assim se $\beta_i \leq 1$ teremos $\| B \|_\infty \leq 1$ e portanto estará satisfeita uma condição suficiente de convergência.

Portanto o método de Gauss-Seidel **converge** se:

a) o **critério de Sassenfeld** for satisfeito, isto é, se:

$$\max_{1 \leq i \leq n} \beta_i < 1$$

onde os β_i são calculados por recorrência através de:

$$\beta_i = \sum_{j=1}^{i-1} |a^*_{ij}| \, \beta_j + \sum_{j=i+1}^{n} |a^*_{ij}| \, . \tag{5.13}$$

b) o **critério das linhas** for satisfeito, isto é, se (5.7) for verificado.

Para provar que o critério das linhas é válido também para o método de Gauss-Seidel, basta verificar que a condição (5.7) implica $\beta_i < 1$, $i = 1, 2, \ldots, n$. De fato, (provando por indução), para $i = 1$, temos:

$$\beta_1 = \sum_{j=2}^{n} |a^*_{1j}| \leq \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a^*_{ij}| < 1.$$

Suponhamos $\beta_j < 1$ para $j = 1, 2, \ldots, i-1$. Segue então:

$$\begin{aligned} \beta_i &= \sum_{j=1}^{i-1} |a^*_{ij}| \, \beta_j + \sum_{j=i+1}^{n} |a^*_{ij}| \\ &\leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a^*_{ij}| \leq \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a^*_{ij}| < 1. \end{aligned}$$

Portanto $\max \beta_i < 1$ e o critério de Sassenfeld é verificado.

c) a matriz dos coeficientes for **estritamente diagonalmente dominante**, isto é, se (5.9) for válido. A prova aqui é idêntica à realizada no método de Jacobi-Richardson.

Observações:

1. Dado um sistema linear $Ax = b$ pode acontecer que o método de Jacobi-Richardson aplicado a ele resulte convergente enquanto que o de Gauss-Seidel resulte divergente e vice-versa.

2. Se $\| B \|$ não for apreciavelmente menor que 1 a convergência pode ser bastante lenta.

3. Uma permutação conveniente das linhas ou colunas de A antes de dividir cada equação pelo coeficiente da diagonal principal pode reduzir o valor de $\parallel B \parallel$.

4. A convergência para os métodos: Jacobi-Richardson e Gauss-Seidel não depende do vetor inicial $x^{(0)}$.

   Evidentemente quanto melhor a aproximação inicial menor será o número de iterações necessárias para atingir uma determinada precisão. Como não conhecemos a priori a solução, normalmente, tomamos o vetor nulo como sendo o vetor inicial. Observe que para o método de Jacobi-Richardson, se tomarmos o vetor nulo teremos $x^{(1)} = b^*$. Tomamos então $x^{(0)} = \theta$ (vetor nulo), para o método de Gauss-Seidel e $x^{(0)} = b^*$, para o método de Jacobi-Richardson.

Exemplo 5.3 - *Resolver o sistema:*

$$\begin{cases} 5x_1 & + & x_2 & + & x_3 & = & 5 \\ 3x_1 & + & 4x_2 & + & x_3 & = & 6 \\ 3x_1 & + & 3x_2 & + & 6x_3 & = & 0 \end{cases} .$$

pelo método de Gauss-Seidel com $\epsilon < 10^{-2}$.

Solução: A matriz dos coeficientes não é estritamente diagonalmente dominante. Assim, por esse critério, nada podemos afirmar sobre a convergência do processo de Gauss-Seidel.

Dividindo cada equação pelo correspondente elemento da diagonal principal obtemos:

$$\begin{cases} x_1 & + & 0.2x_2 & + & 0.2x_3 & = & 1 \\ 0.75x_1 & + & x_2 & + & 0.25x_3 & = & 1.5 \\ 0.5x_1 & + & 0.5x_2 & + & x_3 & = & 0 \end{cases}$$

Vimos anteriormente que se matriz dos coeficientes não for estritamente diagonalmente dominante então o critério das linhas também não será satisfeito. Mas, por se tratar de exemplo, calculemos (5.7). Assim:

$$\begin{aligned} |a^*_{12}| + |a^*_{13}| &= |0.2| + |0.2| \; = \; 0.4 \; , \\ |a^*_{21}| + |a^*_{23}| &= |0.75| + |0.25| \; = \; 1 \; , \\ |a^*_{31}| + |a^*_{32}| &= |0.5| + |0.5| \; = \; 1 \; , \end{aligned}$$

$$\Rightarrow \max_{1 \le i \le n} \sum_{\substack{j=1 \\ j \ne i}}^{3} |a_{ij}| \; = \; 1$$

e portanto por esse critério não podemos garantir convergência.

Aplicando o critério de Sassenfeld, temos:

$$\begin{aligned} \beta_1 &= |0.2| \; + \; |0.2| \; = \; 0.4 \; , \\ \beta_2 &= |0.75|(0.4) \; + \; |0.25| \; = \; 0.3 \; + \; 0.25 \; = \; 0.55 \; , \\ \beta_3 &= |0.5|(0.4) \; + \; |0.5|(0.55) \; = \; 0.2 \; + \; 0.275 \; = \; 0.475 \; , \end{aligned}$$

$$\Rightarrow \max_{1 \le i \le n} \beta_i \; = \; 0.55 < 1 \; ,$$

logo temos o critério de Sassenfeld satisfeito e portanto podemos garantir que o processo de Gauss-Seidel converge .

Temos que as iterações são definidas por:

$$\begin{cases} x_1^{(k+1)} &= -0.2\, x_2^{(k)} \;-\; 0.2x_3^{(k)} \;+\; 1 \\ x_2^{(k+1)} &= -0.75\, x_1^{(k+1)} \;-\; 0.25x_3^{(k)} \;+\; 1.5 \\ x_3^{(k+1)} &= -0.5\, x_1^{(k+1)} \;-\; 0.5x_2^{(k+1)} \end{cases}$$

e a partir de $x^{(0)} = (0,0,0)^t$, obtemos para $x^{(1)}$ os seguintes valores:

$$\begin{cases} x_1^{(1)} &= -0.2x_2^{(0)} - 0.2x_3^{(0)} + 1 = -0.2(0) - 0.2(0) + 1 = 1 \\ x_2^{(1)} &= -0.75x_1^{(1)} - 0.25x_3^{(0)} + 1.5 = -0.75(1) - 0.25(0) + 1.5 = 0.75 \\ x_3^{(1)} &= -0.5x_1^{(1)} - 0.5x_2^{(1)} = -0.5(1) - 0.5(0.75) = -0.875 \end{cases}$$

Continuando as iterações obtemos a tabela:

k	0	1	2	3	4
x_1	0	1	1.025	1.0075	1.0016
x_2	0	0.75	0.95	0.9913	0.9987
x_3	0	- 0.875	- 0.9875	- 0.9994	- 1.0002

Agora, desde que:

$$x^{(4)} - x^{(3)} \;=\; \begin{pmatrix} -0.0057 \\ 0.0074 \\ 0.00075 \end{pmatrix} ,$$

e portanto

$$\frac{\|\, x^{(4)} - x^{(3)}|\, \|_\infty}{\|\, x^{(4)}\, \|_\infty} \;=\; \frac{0.0074}{1.0016} \;\simeq\; 0.0074 \;<\; 10^{-2} ,$$

segue que a solução do sistema, com $\epsilon < 10^{-2}$, é:

$$x \;=\; \begin{pmatrix} 1.0016 \\ 0.9987 \\ -1.0002 \end{pmatrix} .$$

Exercícios

5.3 - *Dado o sistema:*

$$\begin{cases} 4x_1 &+& 2x_2 &+& 6x_3 &=& 1 \\ 4x_1 &-& x_2 &+& 3x_3 &=& 2 \\ -x_1 &+& 5x_2 &+& 3x_3 &=& 3 \end{cases}$$

Mostrar que reordenando as equações e incógnitas podemos fazer com que o critério de Sassenfeld seja satisfeito, mas não o das linhas.

5.4 - *Considere o sistema:*

$$\begin{cases} 5x_1 &+& 2x_2 &+& x_3 &=& 7 \\ -x_1 &+& 4x_2 &+& 2x_3 &=& 3 \\ 2x_1 &-& 3x_2 &+& 10x_3 &=& -1 \end{cases}$$

a) *Verificar a possiblidade de aplicação do método de Gauss-Seidel, usando o critério de Sassenfeld.*

b) *Se possível, resolvê-lo pelo método do item* **a)**, *obtendo o resultado com* $\epsilon < 10^{-2}$.

5.3 Processos de Relaxação

Veremos nessa seção alguns métodos iterativos para resolver sistemas lineares conhecidos como **processos de relaxação**. Para desenvolver tais métodos precisamos de alguns conceitos, os quais passamos a considerar agora.

1) Para uma função $y = f(x)$, o ponto x_0 tal que $f'(x_0) = 0$ é denominado **ponto estacionário** de f. Para saber o tipo de ponto calculamos a derivada segunda de f. Assim se:

   a) $f''(x_0 > 0$ então x_0 é **ponto de mínimo**,

   b) $f''(x_0) < 0$ então x_0 é **ponto de máximo**,

   c) $f''(x_0) = 0$ então x_0 é **ponto de inflexão**.

2) Para uma função de n variáveis $y = f(x_1, x_2, \ldots, x_n)$, denominamos **gradiente** de f, em símbolo, $grad\ f$, :

$$grad\ f = (f_{x_1}, f_{x_2}, \ldots, f_{x_n})\ ,$$

   onde f_{x_i} são as derivadas parciais de f em relação a x_i. Assim o ponto $P = (x_1, \ldots, x_n)^t$ tal que $grad\ f(P) = 0$ é denominado **ponto estacionário** de f.

   Portanto ponto estacionário é o ponto onde todas as derivadas parciais se anulam.

   Para saber o tipo de ponto, devemos calcular as derivadas parciais de 2ª ordem. Seja A uma matriz cujos elementos $(a_{ij}) = \dfrac{\delta^2 f}{\delta x_i \delta x_j}$. Portanto:

$$A(P) = \begin{pmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1 \partial x_2} & \ldots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \ldots & \dfrac{\partial^2 f}{\partial x_2 \partial x_n} \\ \ldots & & & \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1} & \dfrac{\partial^2 f}{\partial x_n \partial x_2} & \ldots & \dfrac{\partial^2 f}{\partial x_n^2} \end{pmatrix} .$$

   Assim, se:

   a) $A(P)$: **positiva definida** então P é **ponto de mínimo**,

   b) $A(P)$: **negativa definida** então P é **ponto de máximo**,

   c) $A(P)$: indefinida então P é **ponto de cela**.

A definição de matriz positiva definida encontra-se no Capítulo 4 (Definição 4.4).
Estamos agora em condições de descrever o processo de relaxação.

Seja o sistema linear :

$$Ax + b = 0\ , \tag{5.14}$$

onde A, $n \times n$, é positiva definida; x e b são vetores $n \times 1$. Portanto, o sistema (5.14) tem uma única solução.

Se v é uma aproximação da solução então:

$$r = Av + b\ ,$$

é o resíduo.

O objetivo do processo de relaxação é fazer com que o resíduo se anule. Para ver como é possível conseguir isso, consideremos, junto com o sistema de equações (5.14), a função quadrática:

$$F(v) = \frac{1}{2}(Av, v) + (b, v) \ , \tag{5.15}$$

onde $A = (a_{ij});\ v = (v_1, v_2, \ldots, v_n)^t; b = (b_1, b_2, \ldots, b_n)^t$, sendo que $(Av, v) \geq 0$, com igualdade válida se e somente se $v = \theta$ (vetor nulo).

Portanto, calculando os produtos escalares da expressão (5.15), obtemos:

$$F(v) = \frac{1}{2} \sum_{i,j=1}^{n} a_{ij} v_i v_j + \sum_{i=1}^{n} b_i v_i \ ,$$

Observe agora que:

$$\begin{aligned}
\sum_{i,j=1}^{n} a_{ij} v_i v_j &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_i v_j \\
&= a_{11} v_1^2 + a_{12} v_1 v_2 + \ldots + a_{1n} v_1 v_n \\
&+ a_{21} v_2 v_1 + a_{22} v_2^2 + \ldots + a_{2n} v_2 v_n \\
&\ \ \ldots \\
&+ a_{n1} v_n v_1 + a_{n2} v_n v_2 + \ldots + a_{nn} v_n^2 \ ,
\end{aligned}$$

$$\sum_{i=1}^{n} b_i v_i = b_1 v_1 + \ldots + b_n v_n \ .$$

Portanto:

$$\frac{\partial \sum_{i,j=1}^{n} a_{ij} v_i v_j}{\partial v_i} = 2 \sum_{j=1}^{n} a_{ij} v_i \ ,$$

desde que A é simétrica, e

$$\frac{\partial \sum_{i=1}^{n} b_i v_i}{\partial v_i} = b_i \ .$$

Logo, podemos escrever que:

$$\begin{aligned}
\frac{\partial F(v)}{\partial v_i} &= \frac{1}{2} \cdot 2 \sum_{j=1}^{n} a_{ij} v_j + b_i \\
&= \sum_{j=1}^{n} a_{ij} v_j + b_i, \ i = 1, \ldots, n \ .
\end{aligned}$$

Agora:

$$grad\ F(v) = \left(\frac{\partial F(v)}{\partial v_1}, \frac{\partial F(v)}{2 v_2}, \ldots \frac{\partial F(v)}{\partial v_n} \right) \ .$$

Portanto:

$$grad\ F(v) = 0 \Leftrightarrow \left(\frac{\partial F(v)}{\partial v_i} \right) = 0 \ , \ i = 1, 2, \ldots, n$$

$$\Leftrightarrow \sum_{j=1}^{n} a_{ij} v_j + b_i \;=\; 0 \;,\; i = 1, \ldots, n \;.$$

Assim, temos que: $Av + b = 0 = grad\ F(v)$. Mas, desde que $Av + b = r$, podemos concluir que $grad\ F(v) = r$. Assim, nosso objetivo é obter $grad\ F(v) = 0$, pois assim teremos $r = 0$.

Teorema 5.2 - *O problema de determinar a solução do sistema (5.14), onde A é positiva definida, é equivalente ao problema de determinar o ponto de mínimo de (5.15) .*

Prova: Evidentemente $P = (x_1, x_2, \ldots, x_n)^t$ é ponto estacionário da F se e somente se $(x_1, x_2, \ldots, x_n)^t$ é solução de (5.14), pois se P é ponto estacionário da F então $grad\ F = 0 \Rightarrow r = 0 \Rightarrow P$ é solução de $Ax + b = 0$. Resta provar que F tem um só ponto estacionário e que este ponto é de mínimo.

Temos que v é ponto estacionário da F se e somente se $grad\ F(v) = 0$, isto é, se e somente se:

$$\sum_{j=1}^{n} a_{ij} v_j + b_i \;=\; 0, \; i = 1, 2, \ldots, n \;.$$

Além disso, o ponto estacionário é único, pois, o sistema admite uma única solução. Como:

$$\frac{\partial^2 F(v)}{\partial v_1^2} \;=\; a_{11} \;,\quad \frac{\partial^2 F(v)}{\partial v_1 v_2} \;=\; a_{12} \;,\quad \ldots \;,\quad \frac{\partial^2 F(v)}{\partial v_i v_j} \;=\; a_{ij} \;,$$

temos que:

$$A \;=\; (a_{ij}) = \frac{\partial^2 F(v)}{\partial v_i v_j} \;.$$

Agora, por hipótese, A é positiva definida. Assim v é ponto de mínimo.

O exemplo a seguir ilustra o teorema anterior.

Exemplo 5.4 - *Seja o sistema linear $Ax + b = 0$, dado por:*

$$\begin{cases} 100\ x_1 + x_2 - 1 & = \quad 0 \\ \\ x_1 + 100 x_2 - 100 & = \quad 0 \end{cases}$$

Calcule a função quadrática dada por (5.15) e mostre que o ponto de mínimo desta função é solução do sistema dado.

Solução: É fácil verificar que a solução do sistema dado é:

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \;.$$

Formemos a função quadrática, $F(v)$. Temos que:

$$Av \;=\; \begin{pmatrix} 100 & 1 \\ 1 & 100 \end{pmatrix} \quad \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \quad \begin{pmatrix} 100 v_1 + v_2 \\ v_1 + 100 v_2 \end{pmatrix} \;.$$

Assim:

$$(Av, v) \;=\; 100 v_1^2 + 2 v_1 v_2 + 100 v_2^2 \;,$$

e

$$(b, v) \;=\; -v_1 - 100v_2 \;.$$

Logo:

$$F(v) \;=\; \frac{1}{2}(100v_1^2 + 2v_1v_2 + 100v_2^2) - v_1 - 100v_2 \;.$$

Além disso:

$$\begin{pmatrix} \dfrac{\partial^2 F}{\partial v_1^2} & \dfrac{\partial^2 F}{\partial v_1 \partial v_2} \\ \\ \dfrac{\partial^2 F}{\partial v_2 \partial v_1} & \dfrac{\partial^2 F}{\partial v_2^2} \end{pmatrix} \;=\; \begin{pmatrix} 100 & 1 \\ \\ 1 & 100 \end{pmatrix} \;=\; A \;,$$

que é uma matriz positiva definida.

Portanto $F(v)$ tem ponto de mínimo em $(0, 1)$, que é a solução do sistema. O valor do mínimo é $F_{min} = -50$. Assim, apesar da forma quadrática ser positiva definida, o mínimo pode ser negativo.

Observe que os métodos de relaxação são usados apenas para sistemas cuja matriz dos coeficientes são positivas definidas. Se A é não singular, mas não é positiva definida não podemos aplicar o raciocínio apresentado. Os métodos de relaxação nestes casos não convergem.

5.3.1 Príncipios Básicos do Processo de Relaxação

Sejam v a solução inicial e $r = Av + b$ o resíduo. Escolhemos uma direção p e variamos v nessa direção, com o objetivo de diminuir $F(v)$, para ir atingindo seu ponto do mínimo que é a solução do sistema; ou seja, tentamos anular o resíduo na direção p.

Assim, variando v na direção p, isto é, tomando:

$$v' = v + tp \;,$$

procuramos determinar o parâmetro t de modo que a função F diminua. Logo devemos procurar o mínimo de F na direção p. Temos então que:

$$\begin{aligned} F(v') &= \frac{1}{2}(Av', v') + (b, v') \\ &= \frac{1}{2}(A(v + tp), v + tp) + (b, v + tp) \\ &= \frac{1}{2}\;\left[(Av, v) + 2t(Av, p) + t^2(Ap, p)\right. \\ &+ \left. 2(b, v) + 2t(b, p)\right] \\ &= F(v) + \frac{t^2}{2}(Ap, p) + t(Av + b, p) \;. \end{aligned}$$

desde que $\frac{1}{2}\;[(Av, v) + 2t(b, v)] = F(v)$. Portanto:

$$F(v') = F(v) + \frac{t^2}{2}(Ap, p) + t(r, p) \;,$$

que é função do parâmetro t.

O parâmetro t é selecionado de tal forma que F é mínimo dentro do conjunto acima examinado.A condição necessária para que isso ocorra é:

$$\begin{aligned} \frac{\partial F(v')}{\partial t} &= t(Ap,p) + (r,p) = 0 \\ \Rightarrow t &= -\frac{(r,p)}{(Ap,p)} \, , \end{aligned}$$

que é um ponto estacionário da F. Além disso $\dfrac{\partial^2 F(v')}{\partial t^2} = (Ap,p) > 0$ pois A é positiva definida e assim t é mínimo na direção p.

Portanto:

$$t_{min} = -\frac{(r,p)}{(Ap,p)} \; . \tag{5.16}$$

Observações:

1) Diferentes escolhas da direção p nos dão diferentes métodos de relaxação.

2) O ponto v', que é tomado na direção p de relaxação com $t = t_{min}$, é chamado **ponto do mínimo**.

3) Analisando a equação (5.16) com $r = Av + b$, notamos que a direção p de relaxação não deve ser escolhida ortogonal ao vetor resíduo r. Se assim fosse, o ponto v' teria sempre $t_{min} = 0$ e não haveria melhoria na aproximação da solução.

Teorema 5.3 - *Para o ponto de mínimo v' com $t = t_{min}$ o novo resíduo $r' = Av' + b$ é ortogonal à direção p da relaxação.*

Prova: Temos que:

$$\begin{aligned} r' &= Av' + b \;=\; A(v + tp) + b \\ &= Av + b + tAp \\ \Rightarrow r' &= r + tAp \; . \end{aligned}$$

Portanto:

$$(r', p) \;=\; (r + tAp, p) \;=\; (r,p) + t(Ap,p) \; .$$

Para $t = t_{min}$, segue que:

$$(r', p) \;=\; (r,p) - \frac{(r,p)}{(Ap,p)}(Ap,p) \;=\; 0 \; .$$

Logo r' e p são ortogonais, demonstrando o teorema. Observe que o novo resíduo não tem componentes na direção p, ou seja, o novo resíduo, r', se anula na direção p.

5.3.2 Método dos Gradientes

Dado o sistema (5.14), com A positiva definida, construímos a função quadrática $F(v)$. Vimos que a solução do sistema dado coincide com o ponto de mímimo de $F(v)$ e que $gradF(v) = Av + b = r$. Aqui, definiremos a direção p de relaxação por:

$$p^k = -r^{(k-1)} \quad \text{para} \quad k = 1, 2, \ldots \tag{5.17}$$

Esta direção é dirigida para o ponto do mínimo.

Todo processo iterativo onde a direção p de relaxação é a do resíduo em sentido oposto é chamado **Método dos Gradientes**.

Temos que:

$$\begin{aligned} t_{min} &= -\frac{(r,p)}{(Ap,p)} = -\frac{(r^{k-1}, p^{(k)})}{(Ap^{(k)}, p^{(k)})} \\ &= \frac{(r^{(k-1)}, r^{(k-1)})}{(Ar^{(k-1)}, r^{(k-1)})} \ , \end{aligned}$$

usando (5.17).

Vimos que ao usar t_{min}, o novo resíduo é ortogonal à direção de relaxação. Portanto neste processo **resíduos consecutivos são ortogonais**, isto é:

$$(r^{(k)}, r^{k-1}) = 0 \quad , \quad k = 1, 2, \ldots$$

Assim, no método dos gradientes, temos que:

$$\begin{aligned} v^{(k)} &= v^{(k-1)} + t\ p^{(k)} \\ \Rightarrow\ v^{(k)} &= v^{(k-1)} - t_{min}\ r^{(k-1)} \ , \end{aligned} \tag{5.18}$$

e ,

$$\begin{aligned} r^{(k)} &= Av^{(k)} + b \\ &= A(v^{(k-1)} - t_{min}\ r^{(k-1)}) + b \\ &= Av^{(k-1)} + b - t_{min}\ Ar^{(k-1)} \\ \Rightarrow r^{(k)} &= r^{(k-1)} - t_{min}\ Ar^{(k-1)} \ . \end{aligned} \tag{5.19}$$

Resumindo: dados $v^{(0)}$ e ϵ, onde ϵ é uma precisão pré-fixada, para aplicar o método dos gradientes, devemos calcular:

a) $\mathrm{r}^{(0)} = \mathrm{Av}^{(0)} + \mathrm{b}$

b) para k = 1, 2, . . .

$$\mathbf{b.1)} \quad \mathrm{t_{min}} = \frac{(\mathrm{r^{(k-1)}}, \mathrm{r^{(k-1)}})}{(\mathrm{Ar^{(k-1)}}, \mathrm{r^{(k-1)}})}$$

$$\mathbf{b.2)} \quad \mathrm{v^{(k)}} = \mathrm{v^{(k-1)}} - \mathrm{t_{min} r^{(k-1)}}$$

$$\mathbf{b.3)} \quad \mathrm{r^{(k)}} = \mathrm{r^{(k-1)}} - \mathrm{t_{min} Ap^{(k-1)}}$$

$$\mathbf{b.4)} \quad \text{Se } \| \mathrm{r^{(k)}} \| < \epsilon, \text{ ouse } \frac{\| \mathrm{v^{(k+1)}} - \mathrm{v^{(k)}} \|}{\| \mathrm{v^{(k+1)}} \|} < \epsilon, \text{ Fim,}$$

caso contrario **b**) .

Exemplo 5.5 - *Usando o método dos gradientes obter a solução do sistema:*

$$\begin{pmatrix} 10 & 1 & 0 \\ 1 & 10 & 1 \\ 0 & 1 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} 11 \\ 11 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

com precisão de 10^{-1}.

Solução: Seja $v^{(0)} = (0,0,0)^t$, então:

$$r^{(0)} = Av^{(0)} + b = \begin{pmatrix} -11 \\ -11 \\ -1 \end{pmatrix},$$

Para $k = 1$, obtemos:

$$(r^{(0)}, r^{(0)}) = 121 + 121 + 1 = 243 ,$$

$$Ar^{(0)} = \begin{pmatrix} 10 & 1 & 0 \\ 1 & 10 & 1 \\ 0 & 1 & 10 \end{pmatrix} \begin{pmatrix} -11 \\ -11 \\ -1 \end{pmatrix} = \begin{pmatrix} -121 \\ -122 \\ -21 \end{pmatrix},$$

$$(Ar^{(0)}, r^{(0)}) = 1331 + 1342 + 21 = 2694 .$$

Assim:

$$t_{min} = \frac{243}{2694} = 0.0902 .$$

Agora,

$$\begin{aligned} v^{(1)} &= v^{(0)} - t_{min}\, r^{(0)} \\ &= \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} - 0.0902 \begin{pmatrix} -11 \\ -11 \\ -1 \end{pmatrix} = \begin{pmatrix} 0.9922 \\ 0.9922 \\ 0.0902 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned} r^{(1)} &= r^{(0)} - t_{min}\, Ar^{(0)} \\ &= \begin{pmatrix} -11 \\ -11 \\ -1 \end{pmatrix} - 0.0902 \begin{pmatrix} -121 \\ -122 \\ -21 \end{pmatrix} = \begin{pmatrix} -0.0858 \\ 0.0044 \\ 0.8942 \end{pmatrix}. \end{aligned}$$

De maneira análoga, para $k = 2$, obtemos:

$$(r^{(1)}, r^{(1)}) = 0.8070 ,$$

$$Ar^{(1)} = \begin{pmatrix} -0.5360 \\ 0.8524 \\ 8.9284 \end{pmatrix} ,$$

$$(Ar^{(1)}, r^{(1)}) = 8.0336 .$$

Assim:

$$t_{min} = 0.1005 .$$

Portanto,

$$v^{(2)} = \begin{pmatrix} 1.0008 \\ 0.9918 \\ 0.0003 \end{pmatrix} ,$$

Agora, desde que:

$$\frac{\| v^{(k+1)} - v^{(k)} \|_\infty}{\| v^{(k+1)} \|_\infty} = \frac{0.0899}{1.0008} \simeq 0.09 < 10^{-1} ,$$

temos que $v^{(2)}$ é solução do sistema dado, com $\epsilon < 10^{-1}$.

Exercícios

5.5 - *Deseja-se resolver um sistema $Ax + b = 0$, onde a é real e :*

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & a \end{pmatrix} .$$

pelo método dos gradientes.

a) *Quais os valores possíveis para a?*

b) *Sendo $b = (1,\ 2,\ 3)^t$ e considerando $a = 0.4$, obtenha a solução do sistema com duas casas decimais corretas usando o método dos gradientes.*

5.6 - *Usando o método dos gradientes, obtenha a solução do sistema:*

$$\begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{pmatrix} 5 \\ 4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} ,$$

com erro relativo inferior a 10^{-3}.

5.3.3 Método dos Gradientes Conjugados

Um outro método de relaxação é o chamado **Método dos Gradientes Conjugados**.

Definição 5.3 - *Dada a aplicação linear A; positiva definida, x e y são direções conjugadas se*

$$(Ax, y) \;=\; (x, Ay) = 0 \; .$$

O primeiro passo no método dos gradientes conjugados é igual ao primeiro passo do método dos gradientes; isto é, dado $v^{(0)}$, calculamos $r^{(0)} = Av^{(0)} + b$ e fazemos:

$$p^{(1)} \;=\; -r^{(0)} \; ,$$

$$v^{(1)} \;=\; v^{(0)} - tr^{(0)} \; ,$$

onde

$$t \;=\; q_1 \;=\; -\frac{(r^{(0)}, p^{(1)})}{(Ap^{(1)}, p^{(1)})} \;=\; \frac{(r^{(0)}, r^{(0)})}{(Ar^{(0)}, r^{(0)})} \; .$$

Portanto:

$$v^{(1)} = v^{(0)} - \frac{(r^{(0)}, r^{(0)})}{(Ar^{(0)}, r^{(0)})}\; r^{(0)} \; . \tag{5.20}$$

Consideremos a passagem do passo $\mathbf{k-1}$ para o passo $\mathbf{k}$, $(k \geq 1)$.

Tomamos a direção de liberação $p^{(k)}$ de tal modo que $p^{(k)}$ e $p^{(k-1)}$ sejam direções conjugadas, isto é, $p^{(k)}$ deve ser tal que:

$$(Ap^{(k)}, p^{(k-1)}) \;=\; (p^{(k)}, Ap^{(k-1)}) \;=\; 0 \; .$$

Além disso, $p^{(k)}$ é tomado como combinação linear de $r^{(k-1)}$ e $p^{(k-1)}$, e desde que o coeficiente de $r^{(k-1)}$ é não nulo podemos tomá-lo igual -1.

Portanto:

$$p^{(k)} = -r^{(k-1)} + \alpha_{k-1}p^{(k-1)}, k = 2, 3, \ldots \tag{5.21}$$

onde α_{k-1} é um coeficiente a ser determinado.

Temos que, para $k = 2, 3, \ldots$:

$$(p^{(k)}, Ap^{(k-1)}) = 0$$

$$\Rightarrow \quad (-r^{(k-1)} + \alpha_{k-1}p^{(k-1)}, Ap^{(k-1)}) = 0$$

$$\Rightarrow \quad (-r^{(k-1)}, Ap^{(k-1)}) + \alpha_{k-1}(p^{(k-1)}, Ap^{(k-1)}) = 0 \; .$$

Da expressão acima podemos determinar α_{k-1}, isto é:

$$\alpha_{k-1} \;=\; \frac{(r^{(k-1)}, Ap^{(k-1)})}{(p^{(k-1)}, Ap^{(k-1)})} \; , \; k = 2, 3, \ldots \; . \tag{5.22}$$

Uma vez, identificada a direção $p^{(k)}$, procuramos o ponto de mínimo. Assim, de $v^{(k)} = v^{(k-1)} + tp^{(k)}$, obtemos que:

$$v^k \;=\; v^{(k-1)} + q_k p^{(k)} \; , \tag{5.23}$$

onde

$$q_k = -\frac{(r^{(k-1)}, p^{(k)})}{(Ap^{(k)}, p^{(k)})} \; .$$

e de $r^{(k)} = Av^{(k)} + b$, segue que:

$$\begin{aligned} r^{(k)} &= A(v^{(k-1)} + q_k p^{(k)}) + b \\ &= Av^{(k-1)} + b + q_k Ap^{(k)} \; . \end{aligned}$$

Portanto:

$$r^{(k)} = r^{(k-1)} + q_k Ap^{(k)} \; . \tag{5.24}$$

Observações:

i) O método dos gradientes conjugados é essencialmente definido pelas fórmulas (5.20), (5.21), (5.22), (5.23), (5.24), desde que um vetor aproximação $v^{(0)}$ tenha sido escolhido.

ii) Os denominadores que aparecem nas fórmulas de α_{k-1} e q_k são sempre maiores que zero, para direções não nulas $p^{(k)}$, pelo fato de A ser positiva definida.

iii) O **resíduo em cada passo**, do método dos gradientes conjugados, possui as seguintes propriedades:

1) é ortogonal ao resíduo do passo anterior, isto é:

$$(r^{(k)}, r^{(k-1)}) = 0 \; ,$$

2) é ortogonal à direção de relaxação do passo, isto é:

$$(r^{(k)}, p^{(k)}) = 0 \; ,$$

3) é ortogonal à direção de relaxação do passo anterior, isto é:

$$(r^{(k)}, p^{(k-1)}) = 0 \; .$$

Com estas propriedades, podemos obter simplificações para as fórmulas de q_k e de α_{k-1}. De fato, da fórmula de q_k, obtemos que:

$$q_k = \frac{(r^{(k-1)}, r^{(k-1)})}{(Ap^{(k)}, p^{(k)})} \; ,$$

desde que:

$$\begin{aligned} -(r^{(k-1)}, p^{(k)}) &= -(r^{(k-1)}, -r^{(k-1)} + \alpha_{k-1} p^{(k-1)}) \\ &= (r^{(k-1)}, r^{(k-1)}) - \alpha_{k-1}(r^{(k-1)}, p^{(k-1)}) \\ &= (r^{(k-1)}, r^{(k-1)}) \; , \; \text{(usando a propriedade 2)} \; , \end{aligned}$$

e da fórmula de α_{k-1}, segue que:

$$\alpha_{k-1} = \frac{(r^{(k-1)}, r^{(k-1)})}{(r^{(k-2)}, r^{(k-2)})} \; ,$$

desde que, de (5.24), temos:

$$Ap^{(k-1)} = \frac{1}{q_{k-1}}(r^{(k-1)} - r^{(k-2)}) \; .$$

Assim:

$$\begin{aligned}
(r^{(k-1)}, Ap^{(k-1)}) &= (r^{(k-1)}, \frac{1}{q_{k-1}}(r^{(k-1)} - r^{(k-2)})) \\
&= \frac{1}{q_{k-1}}(r^{(k-1)}, r^{(k-1)}) - \frac{1}{q_{k-1}}(r^{(k-1)} - r^{(k-2)}) \\
&= \frac{1}{q_{k-1}}(r^{(k-1)}, r^{(k-1)}) \; , \; \text{(usando a propriedade 1)} \; ,
\end{aligned}$$

e

$$\begin{aligned}
(p^{(k-1)}, Ap^{(k-1)}) &= \frac{1}{q_{k-1}}(p^{(k-1)}, r^{(k-1)}) - \frac{1}{q_{k-1}}(p^{(k-1)}, r^{(k-2)}) \\
&= -\frac{1}{q_{k-1}}(p^{(k-1)}, r^{(k-2)}) \; , \text{(usando a propriedade 2)} \\
&= -\frac{1}{q_{k-1}}(-r^{(k-2)} + \alpha_{k-2}p^{(k-2)}, r^{(k-2)}) \\
&= \frac{\alpha_{k-2}}{q_{k-1}}(r^{(k-2)}, p^{(k-2)}) + \frac{1}{q_{k-1}}(r^{(k-2)}, r^{(k-2)}) \\
&= \frac{1}{q_{k-1}}(r^{(k-2)}, r^{(k-2)}) \; , \text{usandoapropriedade2} \; .
\end{aligned}$$

Resumindo, para aplicarmos o método dos gradientes conjugados devemos efetuar os seguintes passos:

Dado $v^{(0)}$ e ϵ, calcular:

a)

$$\begin{aligned}
\mathrm{r}^{(0)} &= \mathrm{Av}^{(0)} + \mathrm{b} \\
p^{(1)} &= -\, r^{(0)} \\
q_1 &= \frac{(r^{(0)}, r^{(0)})}{(Ar^{(0)}, r^{(0)})} \\
v^{(1)} &= v^{(0)} + q_1 p^{(1)} \\
r^{(1)} &= r^{(0)} + q_1 Ap^{(1)}
\end{aligned}$$

b) para $\mathrm{k} \geq 2$

$$\mathbf{b.1)} \quad \alpha_{\mathrm{k-1}} = \frac{(\mathrm{r}^{(\mathrm{k-1})}, \mathrm{r}^{(\mathrm{k-1})})}{(\mathrm{r}^{(\mathrm{k-2})}, \mathrm{r}^{(\mathrm{k-2})})}$$

$$\mathbf{b.2)} \quad \mathrm{p}^{(\mathrm{k})} = -\, \mathrm{r}^{(\mathrm{k-1})} + \alpha_{\mathrm{k-1}} \mathrm{p}^{(\mathrm{k-1})}$$

$$\mathbf{b.3)} \quad \mathrm{q}_{\mathrm{k}} = \frac{(\mathrm{r}^{(\mathrm{k-1})}, \mathrm{r}^{(\mathrm{k-1})})}{(\mathrm{Ap}^{(\mathrm{k})}, \mathrm{p}^{(\mathrm{k})})}$$

$$\mathbf{b.4)} \quad \mathrm{v}^{(\mathrm{k})} = \mathrm{v}^{(\mathrm{k-1})} + \mathrm{q}_{\mathrm{k}} \mathrm{p}^{(\mathrm{k})}$$

$$\mathbf{b.5)} \quad \mathrm{r}^{(\mathrm{k})} = \mathrm{r}^{(\mathrm{k-1})} + \mathrm{q}_{\mathrm{k}} \mathrm{Ap}^{(\mathrm{k})}$$

$$\mathbf{c)} \quad \text{Se } \frac{\| \mathrm{v}^{(\mathrm{k+1})} - \mathrm{v}^{(\mathrm{k})} \|}{\| \mathrm{v}^{(\mathrm{k+1})} \|} < \epsilon, \text{ Fim}$$

caso contrário **b**) .

Teorema 5.4 - *No método dos gradientes conjugados, as direções de relaxação formam um sistema de direções conjugadas e os resíduos formam um sistema ortogonal, isto é, para $i \neq j$, $i, j = 1, 2, \ldots$, vale que:*

$$(Ap^{(i)}, p^{(j)}) = 0 \,,$$

$$(r^{(i)}, r^{(j)}) = 0 \,,$$

Prova: A prova deste teorema pode ser encontrada em Rutishauser, Stiefel, Schwarz.

Do fato de $(r^{(i)}, r^{(j)}) = 0$, concluímos que o método dos gradientes conjugados converge, teoricamente, em n passos, onde n é a ordem do sistema; isto porque os vetores $r^{(i)}$ pertencem a um espaço vetorial n-dimensional, e assim o sistema ortogonal pode conter no máximo n vetores não nulos. Com isso podemos enunciar o seguinte:

Teorema 5.5 - *O método dos gradientes conjugados fornece a solução do sistema em no máximo n passos, onde n é a ordem do sistema.*

Observe que em geral, na prática, devido aos erros de arredondamento, não obteremos a solução do sistema em n passos.

Exemplo 5.6 - *Usando o método dos gradientes conjugados, obter a solução do sistema do exemplo anterior, com duas casas decimais corretas .*

Solução: Como o 1º passo do método dos gradientes conjugados é igual ao 1º passo do método dos gradientes, do exemplo anterior, temos:

$$v^{(0)} = (0, 0, 0)^k \qquad r^{(0)} = \begin{pmatrix} -11 \\ -11 \\ -1 \end{pmatrix} \,,$$

$$p^{(1)} = -r^{(0)} \ , \quad q_1 = t_{min} = 0.09024 \ ,$$

$$v^{(1)} = \begin{pmatrix} 0.9922 \\ 0.9922 \\ 0.0902 \end{pmatrix} , \ r^{(1)} = \begin{pmatrix} -0.0858 \\ 0.0044 \\ 0.8942 \end{pmatrix} .$$

Para $k = 2$, temos:

$$\alpha_1 = \frac{(r^{(1)}, r^{(1)})}{(r^{(0)}, r^{(0)})} = \frac{0.8070}{243} = 0.0033 \ ,$$

$$\begin{aligned} p^{(2)} &= -r^{(1)} + \alpha_1 p^{(1)} \\ &= \begin{pmatrix} 0.0858 \\ -0.0044 \\ -0.8942 \end{pmatrix} + 0.0033 \begin{pmatrix} 11 \\ 11 \\ 1 \end{pmatrix} \\ \Rightarrow p^2 &= \begin{pmatrix} 0.1221 \\ 0.0319 \\ -0.8909 \end{pmatrix} , \end{aligned}$$

$$Ap^{(2)} = \begin{pmatrix} 10 & 1 & 0 \\ 1 & 10 & 1 \\ 0 & 1 & 10 \end{pmatrix} \begin{pmatrix} 0.1221 \\ 0.0319 \\ -0.8909 \end{pmatrix} = \begin{pmatrix} 1.2529 \\ -0.4498 \\ -8.8771 \end{pmatrix} ,$$

$$(Ap^{(2)}, p^{(2)}) = 0.1530 - 0.0143 + 7.9086 = 8.0473 \ ,$$

$$q_2 = \frac{(r^{(1)}, r^{(1)}}{(Ap^{(2)}, p^{(2)})} = \frac{0.8070}{8.0473} = 0.1003 \ ,$$

$$\begin{aligned} v^{(2)} &= v^{(1)} + q_2 p^{(2)} \\ &= \begin{pmatrix} 0.9922 \\ 0.9922 \\ 0.0902 \end{pmatrix} + 0.1003 \begin{pmatrix} 0.1221 \\ 0.0319 \\ -0.8909 \end{pmatrix} = \begin{pmatrix} 1.0044 \\ 0.9954 \\ 0.0008 \end{pmatrix} . \end{aligned}$$

Para $k = 3$, segue que:

$$r^{(2)} = r^{(1)} + q_2 Ap^{(2)} = \begin{pmatrix} 0.0399 \\ -0.0407 \\ 0.0038 \end{pmatrix} ,$$

$$\alpha_2 = \frac{(r^{(2)}, r^{(2)})}{(r^{(1)}, r^{(1)})} = \frac{0.0033}{0.8070} = 0.0041 \ ,$$

$$\begin{aligned} p^{(3)} &= -r^{(2)} + \alpha_2 p^{(2)} \\ &= \begin{pmatrix} -0.0394 \\ 0.0408 \\ -0.0001 \end{pmatrix} , \end{aligned}$$

$$Ap^{(3)} = \begin{pmatrix} -0.3532 \\ 0.3685 \\ 0.0398 \end{pmatrix} ,$$

$$(Ap^{(3)}, p^{(3)}) = 0.0139 + 0.0150 - 0.0000 = 0.0289 ,$$

$$q_3 = \frac{(r^{(2)}, r^{(2)}}{(Ap^{(3)}, p^{(3)})} = \frac{0.0033}{0.0289} = 0.1142 ,$$

$$\begin{aligned} v^{(3)} &= v^{(2)} + q_3 p^{(3)} \\ &= \begin{pmatrix} 0.9999 \\ 1.0001 \\ 0.0008 \end{pmatrix} . \end{aligned}$$

Agora, desde que:

$$\frac{\| v^{(k+1)} - v^{(k)} \|_\infty}{\| v^{(k+1)} \|_\infty} = \frac{0.0047}{1.0001} \simeq 0.0047 < 10^{-2} ,$$

temos que v_3 é solução do sistema dado, com $\epsilon < 10^{-2}$.

Exercícios

5.7 - *Mostre que no método dos gradientes conjugados o resíduo em cada passo é ortogonal ao resíduo anterior, à direção de relaxação do passo e à direção de relaxação do passo anterior.*

5.8 - *Usando o método dos gradientes conjugados resolver os sistemas dados nos exercícios 5.5 e 5.6.*

5.4 Exercícios Complementares

5.9 - *Supomos que o sistema:*

$$\begin{cases} x_1 - \alpha x_2 = c_1 \\ -\alpha x_1 + x_2 - \alpha x_3 = c_2 \\ - \alpha x_2 + x_3 = c_3 \end{cases}$$

seja resolvido iterativamente pelas fórmulas:

$$\begin{cases} x_1^{(k+1)} = \alpha x_2^{(k)} + c_1 \\ x_2^{(k+1)} = \alpha (x_1^{(k)} + x_3^{(k)}) + c_2 \\ x_3^{(k+1)} = \alpha x_2^{(k)} + c_3 \end{cases}$$

Para que valores de α a convergência do método definido acima é garantida? Justifique.

5.10 - *Considere o sistema $Ax = b$; onde:*

$$A = \begin{pmatrix} 10 & -1 & 4 \\ 1 & 10 & 9 \\ 2 & -3 & -10 \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}; \quad b = \begin{pmatrix} 5 \\ 2 \\ 9 \end{pmatrix} .$$

Entre os métodos iterativos que você conhece qual você aplicaria? Por que? Resolva - o pelo método escolhido.

5.11 - *Considere o sistema $Ax = b$; onde:*

$$A = \begin{pmatrix} 50 & -1 & 4 \\ 1 & 50 & 9 \\ 2 & -3 & -50 \end{pmatrix}; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}; \quad b = \begin{pmatrix} 45 \\ 42 \\ 49 \end{pmatrix} .$$

Aplique a este sistema o mesmo método aplicado no exercício anterior. Como se comparam as taxas de convergência? Por que?

5.12 - *Considere os sistemas:*

$$(I) \begin{cases} 5x_1 + 2x_2 + x_3 = 0 \\ 2x_1 + 4x_2 + x_3 = 2 \\ 2x_1 + 2x_2 + 4x_3 = 1 \end{cases} ; \quad (II) \begin{cases} 5x_1 + 4x_2 + x_3 = 2 \\ 3x_1 + 4x_2 + x_3 = 2 \\ 3x_1 + 3x_2 + 6x_3 = -9 \end{cases}$$

Aplicando os critérios que você conhece qual dos métodos iterativos será seguramente convergente? Justifique.

5.13 - *Considere o sistema:*

$$\begin{cases} -x_1 + 2x_2 - x_3 = 1 \\ 2x_1 - x_2 = 1 \\ - x_2 + 2x_3 - x_4 = 1 \\ - x_3 + x_4 = 1 \end{cases}$$

Reordene as equações convenientemente e aplique o método de Gauss - Seidel com garantia de convergência.

5.14 - *Certos sistemas de equações lineares podem ser convenientemente tratados pelo método iterativo de Gauss - Seidel. Depois de uma simples generalização, o método pode ser também usado para alguns sistemas não lineares. Determinar desse modo uma solução do sistema:*

$$\begin{cases} x - 0.1y^2 + 0.05x^2 = 0.7 \\ y + 0.3x^2 - 0.1xz = 0.5 \\ z + 0.4y^2 + 0.1xz = 1.2 \end{cases}$$

com erro relativo inferior a 10^{-2}.

5.15 - *O sistema :*

$$\begin{cases} ax + by + c = 0 \\ dx + ey + f = 0 \end{cases}$$

pode ser resolvido minimizando a função:

$$F = (ax + by + c)^2 + (dx + ey + f)^2 .$$

Começamos com uma solução aproximada (x_k, y_k) *e construímos a seguinte, primeiro mantendo* $y = y_k$ *e variando* x. *O ponto de mínimo é chamado* x_{k+1}. *A seguir, mantendo* $x = x_{k+1}$ *e variando* y. *O ponto de mínimo é chamado* y_{k+1}. *O processo é repetido iterativamente.*

Aplicar o método descrito acima ao sistema:

$$\begin{cases} 5x + 2y - 11 = 0 \\ x - 3y - 9 = 0 \end{cases}$$

5.16 - *Um processo iterativo para resolver sistemas de equações do tipo* $Ax - b = 0$ *é assim definido:*

- *somar* Ix *a ambos os membros, obtendo* $(I + A)x - b = x$,
- *realizar iterações a partir de* $x^{(0)}$ *fazendo:*

$$x^{(k+1)} = (I + A)x^{(k)} - b \ .$$

a) Dê uma condição suficiente que assegure a convergência deste processo iterativo.

b) Aplique este processo para determinar a solução do seguinte sistema:

$$\begin{cases} -1.1x_1 + 0.1x_2 = 1 \\ 0.3x_1 - 0.3x_2 = 0 \end{cases}$$

5.17 - *Considere cada um dos seguintes sistemas de 3 equações:*

$$(I)\begin{cases} 3x_1 - 3x_2 + 7x_3 = 18 \\ x_1 + 6x_2 - x_3 = 10 \\ 10x_1 - 2x_2 + 7x_3 = 27 \end{cases} ; \quad (II)\begin{cases} x_1 + 2x_2 + 5x_3 = 20 \\ x_1 + 3x_2 + x_3 = 10 \\ 4x_1 + x_2 + 2x_3 = 12 \end{cases}$$

a) Sem rearranjar as equações, tente achar as soluções iterativamente, usando os métodos de Jacobi e de Gauss - Seidel, começando com $x^{(0)} = (1.01, 2.01, 3.01)^t$.

b) Rearranje as equações de tal modo que satisfaçam os critérios de convergência e repita o que foi feito no item **a)**.

c) Verifique suas soluções nas equações originais.

5.18 - *Considere o sistema:*

$$\begin{pmatrix} -1 & 2 & -1 & 0 \\ 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 9 \\ 11 \end{pmatrix} .$$

a) *É possível aplicar a este sistema os métodos iterativos que você conhece com garantia de convergência?*

b) *Reordene as equações convenientemente de tal forma que seja possível aplicar o método de Gauss-Seidel com garantia de convergência.*

5.19 - *Dado o sistema linear:*

$$\begin{pmatrix} \alpha & 2 & -2 \\ 1 & \alpha & 1 \\ 2 & 2 & \alpha \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} .$$

para que valores de α *haverá convergência se desejarmos utilizar o método de Jacobi?*

5.20 - *Considere o sistema linear do exercício anterior com* $\alpha = 1$ *e* $x^{(0)} = (1,\ 2,\ 3)^t$. *A aplicação do método de Jacobi fornece a tabela:*

k	0	1	2	3
x_1	1	3	-1	-1
x_2	2	-2	2	2
x_3	3	-3	1	1

Existe alguma contradição com o exercício anterior? Você saberia explicar porque o método de Jacobi convergiu.

5.21 - *Considere o sistema linear* $Ax = b$, *onde:*

$$A = \begin{pmatrix} 20 & 3 & 1 \\ a & 20 & 1 \\ 1 & a & 6 \end{pmatrix} .$$

Para que valores de a *o critério das linhas é verificado?*

5.22 - *Supondo que o sistema linear* $Ax = b$, *onde* A *é a matriz do exercício anterior, esteja sendo resolvido pelo método de Jacobi-Richardson, para quais valores de* a, *pode-se afirmar que:*

$$\| x^{(k)} - \bar{x} \|_\infty \leq \frac{1}{2} \| x^{(k-1)} - \bar{x} \|_\infty ,$$

onde $x^{(k)}$ *e* $x^{(k-1)}$ *são aproximações para a solução e* $\bar{x}$ *é a solução exata.*

5.23 - *O sistema linear* $Ax = b$:

$$(I) \begin{pmatrix} 1 & -a \\ -a & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} , \quad a \in I\!R,$$

pode, sob certas condições, ser resolvido pelo seguinte método iterativo:

$$(II) \begin{pmatrix} 1 & 0 \\ -wa & 1 \end{pmatrix} \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \end{pmatrix} = \begin{pmatrix} 1-w & wa \\ 0 & 1-w \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \end{pmatrix} + \begin{pmatrix} wb_1 \\ wb_2 \end{pmatrix} .$$

a) *Mostre que se* $w = 1$ *o método iterativo* (II) *é o método de Gauss-Seidel.*

b) *Considere em* (I), $a = b_1 = b_2 = 0.5$. *Usando o processo iterativo* (II), *com* $w = 1$, *determine a solução deste sistema com precisão de* $< 10^{-2}$. *Tome como vetor inicial* $x^{(0)} = (0.9, 0.9)^t$.

5.24 - *Dado os sistemas:*

$$(I) \left\{ \begin{array}{rcrcr} 9x_1 & - & x_2 & = & 7 \\ -x_1 & + & 9x_2 & = & 17 \end{array} \right. ; \quad (II) \left\{ \begin{array}{rcrcr} 31x_1 & + & 29x_2 & = & 33 \\ 29x_1 & + & 31x_2 & = & 27 \end{array} \right.$$

a) *Construa as funções quadráticas cujos mínimos são as soluções dos sistemas.*

b) *Determine o número de condição para cada sistema.*

c) *Com base no número de condição de cada sistema o que você pode concluir?*

d) *Resolva o sistema* (II) *pelo método dos gradientes conjugados. Qual é a aproximação ao fim de dois estágios?*

5.25 - *Dado o sistema de equações:*

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ -15 \\ 7 \end{pmatrix} .$$

cuja solução é $x = (2,\ 1,\ -3,\ 2)^t$,

a) *resolva-o pelo método dos gradientes conjugados, efetuando os cálculos com 4 algarismos significativos;*

b) *mostre a ortogonalidade dos vetores resíduos (como verificação dos cálculos efetuados).*

5.5 Problemas Aplicados e Projetos

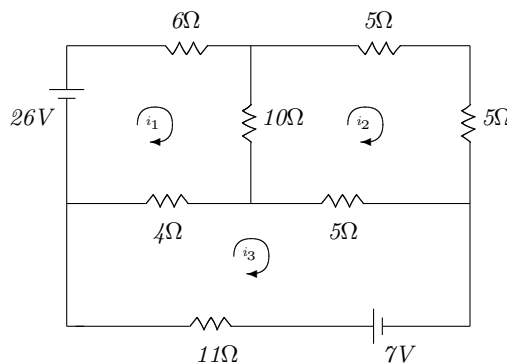
5.1 - *Uma maneira de se obter a solução da equação de Laplace:*

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \ ,$$

em uma região retangular consiste em se fazer uma discretização que transforma a equação em um problema aproximado consistindo em uma equação de diferenças cuja solução, em um caso particular, exige a solução do seguinte sistema linear:

$$\begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 \\ -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 100 \\ 0 \\ 0 \end{pmatrix} .$$

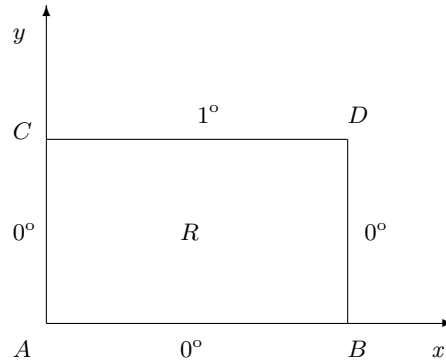
Se desejamos a solução com quatro algarismos significativos corretos, qual dos métodos iterativos que você conhece poderia ser aplicado com garantia de convergência? Resolva o sistema pelo método escolhido.

5.2 - *Considere o circuito da figura a seguir, com resistências e baterias tal como indicado; escolhemos arbitrariamente as orientações das correntes.*

Aplicando a lei de Kirchoff, que diz que a soma algébrica das diferenças de potencial em qualquer circuito fechado é zero, achamos para as correntes i_1, i_2, i_3:

$$\begin{cases} 6i_1 + 10(i_1 - i_2) + 4(i_1 - i_3) - 26 = 0 \\ 5i_2 + 5i_2 + 5(i_2 - i_3) + 10(i_2 - i_1) = 0 \\ 11i_3 + 4(i_3 - i_1) + 5(i_3 - i_2) - 7 = 0 \end{cases}$$

a) *É possível aplicar ao sistema acima o método de Gauss - Seidel com convergência assegurada? Justifique.*

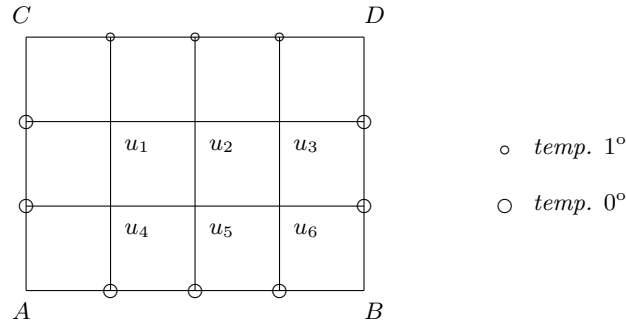
b) *Se possível, obtenha a solução com erro relativo $< 10^{-2}$.*

5.3 - *Suponha uma barra de metal homogêneo, como na figura a seguir, onde $AB = CD = 4 : AC = BD = 3$.*

A temperatura ao longo de AB, AC, BD é mantida constante e igual a 0^oC, enquanto que ao longo de CD ela é igual a 1^oC. A distribuição do calor na barra R obedece à seguinte equação:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \,, \tag{5.25}$$

com as condições de contorno:

$$\begin{aligned} u(x,y) &= 1 \quad \text{para} \quad 0 < x < 4 \,, \\ u(x,0) &= 0 \quad \text{para} \quad 0 < x < 4 \,, \\ u(0,y) &= 1 \quad \text{para} \quad 0 < y < 3 \,, \\ u(x,y) &= 0 \quad \text{para} \quad 0 < y < 3 \,. \end{aligned}$$

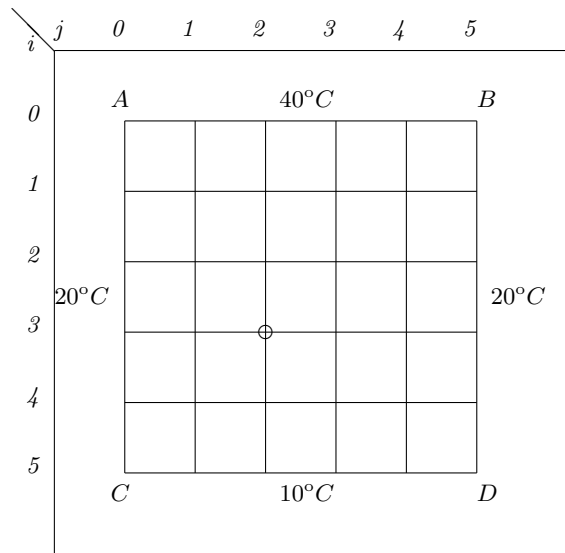
A solução numérica desse problema pode ser obtida considerando-se uma divisão do retângulo $ABCD$ em retângulos menores a partir de uma divisão de AB em intervalos iguais de amplitude h e de uma divisão de CD em intervalos iguais de amplitude k,como é mostrado na figura a seguir:

Nessa figura estamos considerando $h = k = 1$. A temperatura u nos pontos internos pode ser obtida numericamente simulando as derivadas segundas de (5.25), pelas diferenças de segunda ordem $\Delta^2 u$ de modo que para $h = k$, obtemos:

$$\frac{u(x-h,y) - 2u(x,y) + u(x+h,y)}{h^2} + \frac{u(x,y-h) - 2u(x,y) + u(x,y+h)}{h^2} = 0 \ ,$$

para cada par (x,y) em R. Assim, por exemplo, para o ponto $u_1 = u(1,2)$ da figura anterior vale:

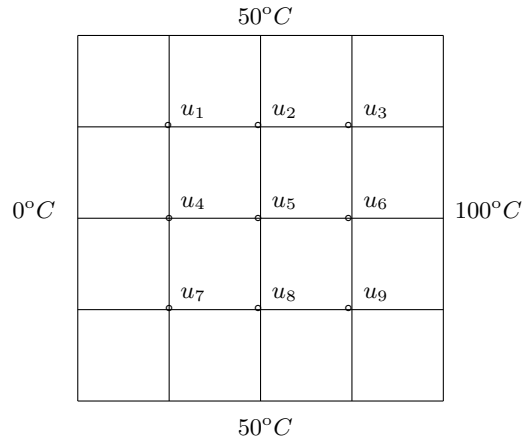
$$\frac{u(0,2) - 2u_1 + u_2}{1^2} + \frac{u_4 - 2u_1 + u(1,3)}{1^2} = 0 \ .$$

Considerando todos os pontos da figura anterior obtemos um sistema de 6 equações lineares nas incógnitas:$u_1, u_2, \ldots, u_6$. Resolva-o por método numérico à sua escolha, com garantia de convergência.

5.4 - *Considere a malha quadrada da figura a seguir, cujos bordos AC e BD são mantidos à temperatura de 20^oC, o bordo AB, à 40^oC, e CD à 10^oC, com o uso de isolantes térmicos em A, B, C, D.*

Para determinar as temperaturas de pontos interiores da malha, pode-se supor que a temperatura em cada ponto é igual à média aritmética dos quatro pontos contíguos. Por exemplo:

$$T_{32} = \frac{T_{22} + T_{31} + T_{33} + T_{42}}{4} \ .$$

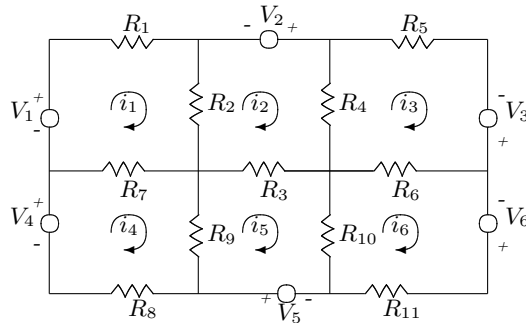
As 16 relações deste tipo permitirão formar um sistema de 16 equações a 16 incógnitas T_{ij}. Resolva-o por método numérico com garantia de convergência.

5.5 - *Suponha que uma membrana com dimensões 80 cm × 80 cm tenha cada um dos seus lados mantidos a uma temperatura constante. Usando a teoria de equações diferenciais parciais pode-se formular uma equação que determina o valor da temperatura no interior dessa membrana. Essa equação diferencial pode ser simplificada colocando-se uma malha com 9 pontos sobre essa membrana e calculando-se a temperatura nos pontos da malha, como mostra a figura:*

onde $u_1, u_2, ..., u_9$ são os valores da temperatura em cada ponto da malha.

O sistema de equações resultante é dado por:

$$\begin{pmatrix} -4 & 1 & & 1 & & & & & \\ 1 & -4 & 1 & & 1 & & & & \\ & 1 & -4 & & & 1 & & & \\ 1 & & & -4 & 1 & & 1 & & \\ & 1 & & 1 & -4 & 1 & & 1 & \\ & & 1 & & 1 & -4 & & & 1 \\ & & & & 1 & & -4 & 1 & \\ & & & & 1 & & 1 & -4 & 1 \\ & & & & & 1 & & 1 & -4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \end{pmatrix} = \begin{pmatrix} -50 \\ -50 \\ -150 \\ 0 \\ 0 \\ -100 \\ -50 \\ -50 \\ -150 \end{pmatrix}.$$

Encontre a solução do sistema por método numérico à sua escolha, com garantia de convergência.

5.6 - *Suponha que tenhamos um circuito que consiste de fontes de tensão independantes e resistores concentrados como mostra a figura:*

A análise completa de tal circuito requer a determinação dos valores das correntes da malha i_k, *indicados na figura, para os valores especificados das fontes de tensão e dos resistores. é necessário, então, formular um sistema de equações simultâneas com as quantidades* i_k *como incógnitas. Cada uma das equações de tal sistema é determinada pela aplicação da lei de Kirchoff em torno das malhas.*

Por exemplo, para a malha definida pela corrente i_1, *temos:*

$$R_1 i_1 + R_2(i_1 - i_2) + R_7(i_1 - i_4) = V_1 \ .$$

Fazendo um cálculo semelhante para as outras malhas obtemos um sistema de 6 equações nas incógnitas $i_1, i_2, i_3, i_4, i_5, i_6$.

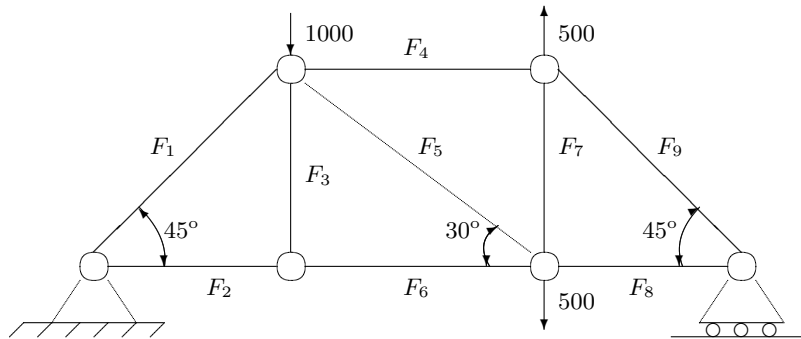
Resolver este problema para os seguintes dados:

$$R = (10, 13, 7, 12, 20, 5, 6, 10, 8, 9, 20)^t \ ,$$

e

$$V = (29, 32, 37, 24, 24, 34)^t \ ,$$

onde R em Ohms e V em Volts.

5.7 - *Numa treliça estaticamente determinada com juntas articuladas, como dada na figura a seguir:*

a tensão, (F_i), *em cada componente pode ser obtida da seguinte equação matricial:*

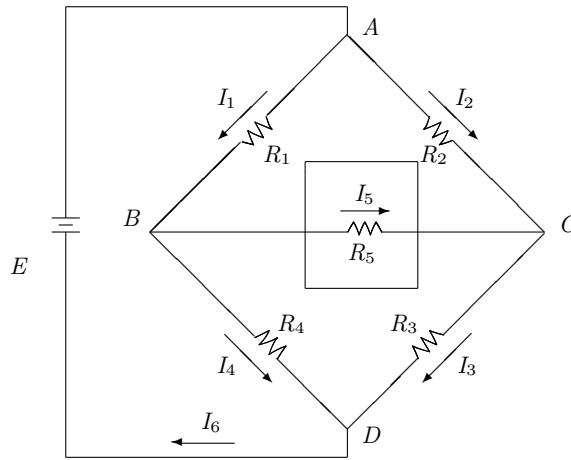
$$\begin{pmatrix} 0.7071 & 0 & 0 & -1 & -0.8660 & 0 & 0 & 0 & 0 \\ 0.7071 & 0 & 1 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.7071 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.7071 \\ 0 & 0 & 0 & 0 & 0.8660 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -0.5 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.7071 \end{pmatrix} F = \begin{pmatrix} 0 \\ -1000 \\ 0 \\ 0 \\ 500 \\ 0 \\ 0 \\ -500 \\ 0 \end{pmatrix} .$$

Observe que as equações são obtidas fazendo-se a soma de todas as forças horizontais ou verticais em cada junta igual a zero. Além disso a matriz dos coeficientes é bastante esparsa, e assim um candidato natural é o método de Gauss-Seidel.

a) *As equações podem ser rearranjadas de modo a se obter uma matriz estritamente diagonalmente dominante?*

b) *É o sistema convergente se iniciarmos com um vetor com todas as componentes iguais a zero?*

c) *Resolva o sistema pelo método de Gauss-Seidel, partindo do vetor nulo e obtendo a solução com precisão de* 10^{-4}.

5.8 - *O circuito mostrado a seguir é frequentemente usado em medidas elétricas e é conhecido com uma "Ponte de Wheatstone".*

As equações que governam o sistema são obtidas a partir da lei de Kirchoff. Para a malha fechada através da bateria e ao longo de ABD, temos:

$$I_1R_1 + I_4R_4 - E = 0 \quad (1)$$

Para a a malha fechada ABCA:

$$I_1R_1 + I_5R_5 - I_2R_2 = 0 \quad (2)$$

Para a malha fechada BCDB:

$$I_5R_5 + I_3R_3 - I_4R_4 = 0 \quad (3)$$

Para o nó A:

$$I_6 = I_1 + I_2 \quad (4)$$

Para o nó B:

$$I_1 = I_5 + I_4 \quad (5)$$

Para o nó C:

$$I_3 = I_2 + I_5 \quad (6)$$

onde R_i *representam as resistências;* I_i *as correntes e* E *a voltagem aplicada.*

Determinar as correntes no problema proposto quando: $E = 20\ Volts$, $R_1 = 10\ Ohms$ *e* $R_2 = R_3 = R_4 = R_5 = 100\ Ohms$.

Capítulo 6

Programação Matemática

Com o objetivo de facilitar o aprendizado do estudante, pretendemos aqui relembrar alguns conceitos básicos, que irão facilitar a compreensão dos métodos numéricos apresentados nos próximos capítulos. É claro que esperamos que o aluno para fazer um curso sobre métodos numéricos já possua conhecimento de alguns conceitos, principalmente, da álgebra linear e do cálculo diferencial e integral. Como é impossível relembrar todos os conceitos dessas duas áreas, salientamos que, a maioria dos conceitos aqui apresentados são de álgebra linear, e uma análise mais profunda pode ser encontrado em livros dessa área.

6.1 Espaço Vetorial

Pretendemos aqui definir importantes noções de dependência linear, base, dimensão e mudança de base em um espaço vetorial.
Este Capítulo é do Marcos Arenales

Capítulo 7

Determinação Numérica de Auto-Valores e Auto-Vetores

7.1 Introdução

Auto-valores e auto-vetores estão presentes em diferentes ramos da matemática incluindo formas quadráticas, sistemas diferenciais; problemas de otimização não linear, e podem ser usados para resolver problemas de diversos campos, como economia, teoria da informação, análise estrutural, eletrônica, teoria de controle e muitos outros.

Nosso objetivo nesse capítulo é apresentar métodos numéricos para a determinação dos auto-valores e correspondentes auto-vetores de uma matriz A de ordem n. Sugerimos ao leitor rever a seção sobre auto-valores e auto-vetores dada no Capítulo 1. A menos que a matriz seja de ordem baixa ou que tenha muitos elementos iguais a zero, a expansão direta do determinante para a determinação do polinômio característico, ver exemplo 1.22, é ineficiente. Assim os métodos numéricos que estudaremos são obtidos sem fazer uso do cálculo do determinante. Tais métodos podem ser divididos em três grupos:

i) métodos que determinam o polinômio característico,

ii) métodos que determinam alguns auto-valores,

iii) métodos que determinam todos os auto-valores.

Nos dois últimos casos determinamos os auto-valores sem conhecer a expressão do polinômio característico.

Em relação aos métodos do grupo **i)**, uma vez determinado o polinômio característico de A, para calcular os auto-valores devemos utilizar métodos numéricos para determinação de zeros de polinômio, (ver Capítulo 3). Nessa classe encontram-se, entre outros, os métodos de Leverrier e Leverrier-Faddeev.

Os métodos do grupo **ii)**, chamados iterativos, são usados se não estamos interessados em todos os auto-valores de A. Incluem-se nessa classe os métodos das potências, potência inversa.

Em relação aos métodos do grupo **iii)**, podemos dividí-los em duas classes:

a) métodos numéricos para matrizes simétricas,

b) métodos numéricos para matrizes não simétricas.

Na classe **a)**, inclui-se entre outros, o método de Jacobi, o qual reduz uma dada matriz simétrica numa forma especial, cujos auto-valores são facilmente determinados. Entre os métodos da classe **b)** podemos citar os métodos de Rutishauser (método LR) e o de Francis (método QR) os quais transformam a

matriz dada numa matriz triangular superior. Todos os métodos do grupo **iii)** fazem uso de uma série de transformações de similaridade e assim são algumas vezes referenciados como métodos de transformações ou métodos diretos.

Maiores detalhes sobre essas técnicas, bem como sobre a teoria desses métodos podem ser encontradas em [Wilkinson,1965].

Descreveremos e exemplificaremos cada um dos métodos numéricos mencionados acima, iniciando com aqueles que determinam o polinômio característico. Antes porém precisamos do seguinte resultado.

Teorema 7.1 - *(Teorema de Newton) - Seja o polinômio:*

$$P(x) \;=\; a_0 x^n \;+\; a_1 x^{n-1} + \;\ldots\; + \;a_{n-1}x \;+\; a_n \;,$$

cujas raízes são: $x_1, x_2, \ldots, x_n$ *. Seja ainda:*

$$s_k \;=\; \sum_{i=1}^{n} x_i^k \quad , \quad 1 \leq k \leq n \;,$$

então:

$$\sum_{i=0}^{k-1} a_i s_{k-1} \;+\; k\; a_k \;=\; 0 \;\;,\;\; k = 1, 2, \ldots, n \;.$$

Prova: A prova deste teorema pode ser encontrada em [Jennings,19..].

Através desse teorema vemos que existe uma relação entre os coeficientes de um polinômio e as somas das potências das suas raízes. Assim, conhecidas as somas das potências das raízes do polinômio podemos determinar os coeficientes do mesmo.

Exemplo 7.1 - *Sejam* $s_1 \;=\; 6,\; s_2 = 14,\; s_3 = 36$ *as somas das potências das raízes de um polinômio* $P(x)$*. Determinar* $P(x)$*.*

Solução: Pelo teorema 7.1, temos:

$$\begin{array}{lll} k = 1 & \Rightarrow & a_0 s_1 \;+\; a_1 \;=\; 0 \;\Rightarrow\; a_1 = -a_0 s_1 \\ k = 2 & \Rightarrow & a_0 s_2 \;+\; a_1 s_1 \;+\; 2a_2 \;=\; 0 \;\Rightarrow\; 2a_2 = -a_0 s_2 - a_1 s_1 \\ k = 3 & \Rightarrow & a_0 s_3 \;+\; a_1 s_2 \;+\; a_2 s_1 \;+\; 3a_3 \;=\; 0 \;\Rightarrow \\ & \Rightarrow & 3a_3 \;=\; -a_0 s_3 - a_1 s_2 - a_2 s_1 \end{array}$$

Tomando o coeficiente do termo de maior grau do polinômio igual a 1, isto é, fazendo $a_0 = 1$, obtemos por substituição nas expressões anteriores que:

$$a_1 \;=\; -6 \;,\;\; a_2 \;=\; 11 \;,\;\; a_3 \;= 6 \;.$$

Portanto, o polinômio procurado é:

$$P(x) \;=\; x^3 - 6x^2 + 11x - 6 \;.$$

Logo, o conhecimento dos $s_k, k = 1, \ldots, n$, proporciona a determinação dos $a_k, k = 1, 2, \ldots, n$. Observe que nesse exemplo as raízes do polinômio são: $x_1 = 1,\; x_2 = 2$ e $x_3 = 3$.

Para os métodos numéricos descritos a seguir usaremos a seguinte notação para o polinômio característico de uma matriz A, de ordem n:

$$P(\lambda) \;=\; (-1)^n \left[\lambda^n - p_1 \lambda^{n-1} - p_2 \lambda^{n-2} - \ldots - p_{n-1}\lambda - p_n\right] \;. \tag{7.1}$$

7.2 Método de Leverrier

O **Método de Leverrier** fornece o polinômio característico de uma matriz A de ordem n.

Seja A uma matriz quadrada de ordem n. Se $\lambda_1, \lambda_2, \ldots, \lambda_n$ são os auto-valores da matriz A, isto é, se $\lambda_1, \lambda_2, \ldots \lambda_n$ são os zeros do polinômio (7.1) e se

$$s_k = \sum_{i=1}^{n} \lambda_i^k \quad , \quad 1 \leq k \leq n ,$$

então, pelo Teorema 7.1, temos:

$$kp_k = s_k - p_1 s_{k-1} - \ldots - p_{k-1} s_1 \quad , \quad 1 \leq k \leq n . \tag{7.2}$$

Portanto, se conhecermos os $s_k, 1 \leq k \leq n$, poderemos determinar os coeficientes $p_1, p_2, \ldots, p_n$ de $P(\lambda)$.

Vejamos então como determinar as somas parciais s_k. Fazendo expansão direta do determinante de $A - \lambda I$, o coeficiente de λ^{n-1} em $P(\lambda)$ é $(-1)^{n-1}(a_{11} + a_{22} + \ldots + a_{nn})$. Por outro lado esse mesmo coeficiente em (7.1) é $(-1)^{n-1}p_1$. Logo devemos ter:

$$p_1 = a_{11} + a_{22} + \ldots + a_{nn} .$$

A soma dos elementos da diagonal principal de uma matriz A é conhecida como **traço** de A, cuja notação é $tr(A)$. Além disso, de (7.2), $s_1 = p_1$, e assim:

$$s_1 = tr\ (A) ,$$

isto é, a soma dos auto-valores da matriz A é igual ao traço de A.

Então, desde que os auto-valores de A^k são a kª potência dos auto-valores de A, (ver exercício 1.26), temos:

$$s_k = tr(A^k) .$$

Assim os números $s_1, s_2, \ldots, s_n$ são obtidos através do cálculo das potências de A, e (7.2) pode ser usada para determinar os coeficientes do polinômio característico. Determinando as raízes desse polinômio por qualquer dos métodos numéricos estudados no Capítulo 3, obtemos os auto-valores de A.

Exemplo 7.2 - *Seja:*

$$A = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix} .$$

Determinar seus auto-valores usando o Método de Leverrier.

Solução: Temos:

$$s_1 = tr(A) = 3 ,$$

$$s_2 = tr(A^2) , \ A^2 = A \cdot A = \begin{pmatrix} 2 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & -1 & 2 \end{pmatrix} , \Rightarrow s_2 = 3 ,$$

$$s_3 = tr(A^3) , \ A^3 = A^2 \cdot A = \begin{pmatrix} 2 & 2 & -2 \\ -1 & -1 & 2 \\ -3 & 1 & 0 \end{pmatrix} , \Rightarrow s_3 = -3 .$$

Usando (7.2), obtemos:

$$\begin{aligned} p_1 &= s_1 \Rightarrow p_1 = 1 \ , \\ 2p_2 &= s_2 - p_1 s_1 \Rightarrow p_2 = 2 \ , \\ 3p_3 &= s_3 - p_1 s_2 - p_2 s_1 \Rightarrow p_3 = -2 \ . \end{aligned}$$

De (7.1), segue que:

$$\begin{aligned} P(\lambda) &= (-1)^3(\lambda^3 - p_1 \lambda^2 - p_2\lambda - p_3) \\ &= (-1)^3 (\lambda^3 - \lambda^2 + 2\lambda - 2) \\ &= -\lambda^3 + 2\lambda^2 - 2\lambda + 2 \ . \end{aligned}$$

Para determinar os auto-valores de A basta determinar os zeros de $P(\lambda)$. É fácil verificar que $\lambda = 1$ é uma raiz de $P(\lambda)$. Usando o algoritmo de Briot-Ruffini-Horner, (Capítulo 3), obtemos:

	-1	1	2	-2
1		-1	0	2
	-1	0	2	0

Assim, $P(\lambda) = (\lambda - 1)(-\lambda^2 + 2)$. Logo os auto-valores de A são: $\lambda_1 = 1$, $\lambda_2 = -\sqrt{2}$ e $\lambda_3 = \sqrt{2}$.

Exercícios

7.1 - *Usando o método de Leverrier, determinar o polinômio característico e os auto-valores do operador* $T: \mathbb{R}^3 \rightarrow \mathbb{R}^3$, *definido por:*

$$T(x,y,z) = (2x + y, \ y - z, \ 2y + 4z) \ .$$

7.2 -*Seja:*

$$A = \begin{pmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{pmatrix} \ .$$

Determinar seu polinômio característico e seus auto-valores pelo processo de Leverrier.

7.3 Método de Leverrier-Faddeev

Uma modificação do método de Leverrier, devida a Faddeev, simplifica os cálculos dos coeficientes do polinômio característico e fornece, em alguns casos, os auto-vetores de A. Tal método é conhecido por **Método de Leverrier-Faddeev**.

Para descrever tal método, definimos uma sequência de matrizes:

$$A_1, A_2, \ldots, A_n,$$

do seguinte modo:

$$\begin{aligned} A_1 &= A \ , \quad q_1 = tr A_1 \ , \quad B_1 = A_1 - q_1 I \ ; \\ A_2 &= AB_1 \ , \quad q_2 = \frac{tr A_2}{2} \ , \quad B_2 = A_2 - q_2 I \ , \\ A_3 &= AB_2 \ , \quad q_3 = \frac{tr A_3}{3} \ , \quad B_3 = A_3 - q_3 I \ ; \\ &\vdots \\ A_n &= AB_{n-1} \ , \quad q_n = \frac{tr A_n}{n} \ , \quad B_n = A_n - q_n I \ . \end{aligned} \tag{7.3}$$

Propriedades da sequência: $A_1, A_2, \ldots, A_n$

1ª) Os termos q_k obtidos na sequência (7.3), são os coeficientes do polinômio característico (7.1), isto é:

$$q_k = p_k,\ k = 1, 2, \ldots, n\ .$$

Prova: A prova será feita por indução.

a) Desde que $A = A_1$, segue que: $q_1 = tr(A_1) = tr(A) = p_1$.

b) Suponhamos que: $q_i = p_i, i = 1, 2, \ldots, k-1$.

c) Provemos que: $q_k = p_k$. Por (7.3), temos:

$$\begin{aligned} A_1 &= A\ , \\ A_2 &= AB_1 = A\,(A_1 - q_1 I) = A\,(A - q_1 I) = A^2 - q_1 A\ , \\ A_3 &= AB_2 = A\,(A_2 - q_2 I) = A\left(A^2 - q_1 A - q_2 I\right) \\ &= A^3 - q_1 A^2 - q_2 A\ , \\ &\vdots \\ A_k &= AB_{k-1} = A\,(A_{k-1} - q_{k-1} I) \\ &= A^k - q_1 A^{k-1} - q_2 A^{k-2} - \ldots - q_{k-1} A\ . \end{aligned}$$

Desde que $q_i = p_i, i = 1, 2, \ldots, k-1$, (hipótese de indução), obtemos:

$$A_k = A^k - p_1 A^{k-1} - p_2 A^{k-2} - \ldots - p_{k-1} A\ . \tag{7.4}$$

Aplicando traço em ambos os membros da igualdade (7.4), segue que:

$$tr(A_k) = tr\left(A^k\right) - p_1 tr\left(A^{k-1}\right) - p_2 tr\left(A^{k-2}\right) - \ldots - p_{k-1} tr(A)\ .$$

Agora, desde que $s_i = tr(A^i),\ i = 1, 2, \ldots, k$, e, por (7.3) $q_k = \dfrac{tr(A_k)}{k}$, obtemos:

$$kq_k = s_k - p_1 s_{k-1} - p_2 s_{k-2} - \ldots - p_{k-2} s_2 - p_{k-1} s_1\ . \tag{7.5}$$

Comparando (7.5) com (7.2), obtemos:

$$q_k = p_k\ ,$$

o que completa a prova.

2ª) Se A é uma matriz de ordem n, então:

$$B_n = \theta \quad \text{(matriz nula)}\ .$$

Prova: Pelo Teorema de Cayley-Hamilton, (Teorema 1.8), temos:

$$A^n - p_1 A^{n-1} - \ldots - p_{n-1} A - p_n I = \theta\ .$$

Mas, por (7.3), e usando a 1ªpropriedade, segue que:

$$B_n = A_n - p_n I\ .$$

Fazendo $k = n$ em (7.4) e substituindo o valor de A_n, na expressão anterior, obtemos:

$$B_n = A^n - p_1 A^{n-1} - \ldots - p_{n-2} A^2 - p_{n-1} A - p_n I = \theta\ .$$

3ª) Se A é uma matriz não singular, de ordem n, então:

$$A^{-1} = \frac{1}{p_n} B_{n-1} .$$

Prova: De $B_n = \theta$ e $B_n = A_n - p_n I$, temos:

$$A_n = p_n I .$$

Mas, por (7.3),

$$A_n = AB_{n-1} .$$

Logo:

$$AB_{n-1} = p_n I .$$

Se A é não singular então existe A^{-1}. Assim, pré-multiplicando ambos os membros da igualdade anterior por A^{-1}, segue que:

$$A^{-1} = \frac{1}{p_n} B_{n-1} .$$

Observações:

a) Com o método de Leverrier-Faddeev, obtemos o polinômio característico de A. Para determinar seus auto-valores basta determinar os zeros de $P(\lambda)$.

b) Se ao fazer os cálculos B_n resultar numa matriz diferente da matriz nula, você terá cometido erros de cálculo.

c) Como $B_n = \theta$ e como $B_n = A_n - p_n I$ então A_n é uma matriz diagonal com todos os elementos não nulos iguais a p_n.

d) Se A é singular então $p_n = 0$. Nesse caso $\lambda = 0$ é um auto-valor de A.

Cálculo dos Auto-Vetores

Sejam $\lambda_1, \lambda_2, \ldots, \lambda_n$ auto-valores distintos de A. Mostraremos a seguir que cada coluna não nula da matriz:

$$Q_k = \lambda_k^{n-1} I + \lambda_k^{n-2} B_1 + \ldots + \lambda_k B_{n-2} + B_{n-1} , \tag{7.6}$$

é um auto-vetor correspondente ao auto-valor λ_k.

Observações:

1) Em (7.6), $B_i, i = 1, \ldots, n-1$, são as matrizes calculadas para a determinação dos coeficientes do polinômio característico, isto é, são as matrizes obtidas em (7.3), e λ_k é o k-ésimo auto-valor de A.

2) Pode-se provar que Q_k é matriz não nula se os auto-valores de A são distintos.

3) Pode ocorrer que mesmo com λ_i iguais a matriz Q_k não seja nula.

Provemos agora que cada coluna não nula de Q_k é um auto-vetor correspondente ao auto-valor λ_k. Temos:

$$\begin{aligned}
(\lambda_k I - A)\ Q_k &= (\lambda_k I - A)\left(\lambda_k^{n-1} I + \lambda_k^{n-2} B_1 + \ldots + \lambda_k B_{n-2} + B_{n-1}\right) \\
&= \lambda_k^n I + \lambda_k^{n-1}(B_1 - A) + \lambda_k^{n-2}(B_2 - AB_1) + \ldots \\
&+ \lambda_k (B_{n-1} - AB_{n-2}) - AB_{n-1} \\
&= \lambda_k^n I - p_1 \lambda_k^{n-1} I - p_2 \lambda_k^{n-2} I - \ldots - p_{n-1}\lambda_k I - p_n I = \theta ,
\end{aligned}$$

desde que λ_k é auto valor de A e portanto é raiz do polinômio característico. Assim, acabamos de mostrar que:

$$AQ_k = \lambda_k Q_k ,$$

Portanto, construídas as matrizes B_i e determinados todos os auto-valores da matriz A, para obter os auto-vetores correspondentes ao auto-valor λ_k basta calcular a matriz Q_k usando (7.6). Entretanto, observe que se u é alguma coluna não nula de Q_k, então, podemos escrever que:

$$Au = \lambda_k u .$$

isto é, u é auto-vetor de A correspondente ao auto-valor λ_k. Assim, ao invés de determinarmos a matriz Q_k, é muito mais vantajoso calcularmos apenas uma coluna u de Q_k, da seguinte maneira: Fazemos,

$$\begin{aligned} u_0 &= e \\ u_i &= \lambda_k u_{i-1} + b_i , \; i = 1, 2, \ldots, n-1 , \end{aligned} \tag{7.7}$$

onde e é uma coluna adotada da matriz identidade e b_i é sua correspondente coluna da matriz B_i, isto é, se adotamos e como sendo a i-ésima coluna da matriz identidade então $b_1, b_2, \ldots, b_{n-1}$ em (7.7) serão, respectivamente, a i-ésima coluna das matrizes $B_1, B_2, \ldots, B_{n-1}$. Logo, $u = u_{n-1}$ é o auto-vetor correspondente ao auto-valor λ_k. Note que em (7.7), i varia de 1 até $n-1$ pois $B_n = \theta$.

Observe que se calcularmos até u_{n-1} e este resultar no vetor nulo, devemos adotar outra coluna da matriz identidade e refazer os cálculos, pois por definição o auto-vetor é um vetor não nulo.

Exemplo 7.3 - *Considere a matriz dada no exemplo 7.2. Usando o método de Leverrier-Faddeev, determinar:*

- **a)** *seu polinômio característico,*
- **b)** *seus auto-valores e correspondentes auto-vetores,*
- **c)** *sua inversa.*

Solução:

a) Para determinar o polinômio característico devemos construir a sequência A_1, A_2, A_3. Assim,

usando (7.3), obtemos:

$$\begin{aligned}
A_1 &= A = \begin{pmatrix} 1 & 1 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}, \; p_1 = tr(A_1) \Rightarrow p_1 = 1, \\
B_1 &= A_1 - p_1 I \Rightarrow B_1 = \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & -1 \end{pmatrix}, \\
A_2 &= AB_1 \Rightarrow A_2 = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 0 & -2 & 2 \end{pmatrix}, \\
p_2 &= \frac{tr(A_2)}{2} \Rightarrow p_2 = \frac{4}{2} \Rightarrow p_2 = 2, \\
B_2 &= A_2 - p_2 I \Rightarrow B_2 = \begin{pmatrix} -1 & -1 & 1 \\ -1 & -1 & -1 \\ 0 & -2 & 0 \end{pmatrix}, \\
A_3 &= AB_2 \Rightarrow A_3 = \begin{pmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \\
p_3 &= \frac{tr(A_3)}{3} \Rightarrow p_3 = \frac{-6}{3} \Rightarrow p_3 = -2, \\
B_3 &= A_3 - p_3 I \Rightarrow B_3 = \theta .
\end{aligned}$$

Usando (7.1), segue que:

$$\begin{aligned}
P(\lambda) &= (-1)^3(\lambda^3 - p_1\,\lambda^2 - p_2\lambda - p_3) \\
&= (-1)^3\,(\lambda^3 - \lambda^2 + 2\lambda - 2) \\
&= -\lambda^3 + 2\lambda^2 - 2\lambda + 2 .
\end{aligned}$$

Para determinar os auto-valores de A basta determinar os zeros de $P(\lambda)$. Já fizemos esses cálculos no exemplo 7.2, e obtivemos: $\lambda_1 = 1$, $\lambda_2 = -\sqrt{2}$ e $\lambda_3 = \sqrt{2}$.

b) Determinemos agora os auto-vetores correspondentes a esses auto-valores.

b.1) Para $\lambda_1 = 1$, seja $e = (1,\ 0,\ 0)^t$. Assim:

$$\begin{aligned}
& u_0 = e \Rightarrow u_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \\
& u_1 = \lambda_1 u_0 + b_1 \Rightarrow u_1 = 1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \Rightarrow u_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \\
& u_2 = \lambda_1 u_1 + b_2 \Rightarrow u_2 = 1 \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \\ 0 \end{pmatrix} \Rightarrow u_2 = \begin{pmatrix} 0 \\ -1 \\ -1 \end{pmatrix} .
\end{aligned}$$

Logo $u = (0,\ -1,\ -1)^t$ é um auto-vetor correspondente ao auto-valor $\lambda_1 = 1$.

Observe que se adotamos $e = (0,\ 1,\ 0)^t$ obtemos $u_2 = (0,\ -1,\ -1)^t$ que é auto-vetor de A correspondente ao auto-valor $\lambda_1 = 1$; mas se adotamos $e = (0,\ 0,\ 1)^t$ obtemos $u_2 = (0,\ 0,\ 0)^t$ e assim com esse vetor inicial não obtemos uma resposta válida.

b.2) Para $\lambda_2 = -\sqrt{2}$, seja $e = (1,\ 0,\ 0)^t$. Assim,

$$u_0 = e \Rightarrow u_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

$$u_1 = \lambda_2 u_0 + b_1 \Rightarrow u_1 = -\sqrt{2} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \Rightarrow u_1 = \begin{pmatrix} -\sqrt{2} \\ 0 \\ -1 \end{pmatrix},$$

$$u_2 = \lambda_2 u_1 + b_2 \Rightarrow u_2 = -\sqrt{2} \begin{pmatrix} -\sqrt{2} \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \\ 0 \end{pmatrix} \Rightarrow u_2 = \begin{pmatrix} 1 \\ -1 \\ \sqrt{2} \end{pmatrix}.$$

Logo $u = (1,\ -1,\ \sqrt{2})^t$ é um auto-vetor correspondente ao auto-valor $\lambda_2 = -\sqrt{2}$.

Novamente, observe que se adotamos $e = (0,\ 1,\ 0)^t$ obtemos $u_2 = (-1-\sqrt{2},\ 1+\sqrt{2},\ -2-\sqrt{2})^t$, enquanto que $e = (0,\ 0,\ 1)^t$ fornece $u_2 = (1+\sqrt{2},\ -1-\sqrt{2},\ 2+\sqrt{2})^t$. Ambos são auto-vetores de A correspondentes ao auto-valor $\lambda_2 = -\sqrt{2}$.

b.3) Para $\lambda_3 = \sqrt{2}$, seja $e = (1,\ 0,\ 0)^t$. Assim:

$$u_0 = e \Rightarrow u_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix},$$

$$u_1 = \lambda_3 u_0 + b_1 \Rightarrow u_1 = \sqrt{2} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix} \Rightarrow u_1 = \begin{pmatrix} \sqrt{2} \\ 0 \\ -1 \end{pmatrix},$$

$$u_2 = \lambda_3 u_1 + b_2 \Rightarrow u_2 = \sqrt{2} \begin{pmatrix} \sqrt{2} \\ 0 \\ -1 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \\ 0 \end{pmatrix} \Rightarrow u_2 = \begin{pmatrix} 1 \\ -1 \\ -\sqrt{2} \end{pmatrix}.$$

Logo $u = (1,\ -1,\ -\sqrt{2})^t$ é um auto-vetor correspondente ao auto-valor $\lambda_3 = \sqrt{2}$.

Observe que se adotamos $e = (0,\ 1,\ 0)^t$ obtemos $u_2 = (-1+\sqrt{2},\ 1-\sqrt{2},\ -2+\sqrt{2})^t$, enquanto que $e = (0,\ 0,\ 1)^t$ fornece $u_2 = (1-\sqrt{2},\ -1+\sqrt{2},\ 2-\sqrt{2})^t$. Novamente, ambos são auto-vetores de A correspondentes ao auto-valor $\lambda_3 = \sqrt{2}$.

Finalmente observe que para cada auto-valor λ_k, a escolha do vetor inicial produz exatamente a coluna correspondente da matriz Q_k. Entretanto, como pode ser observado nesse exemplo, não é necessário calcular todas as colunas da matriz Q_k, isto é , basta uma, pois as colunas não nulas de Q_k são múltiplas uma das outras.

c) Pela 3ª propriedade, temos:

$$A^{-1} = \frac{1}{p_3} B_2 ,$$

e assim:

$$A^{-1} = \frac{1}{-2} \begin{pmatrix} -1 & -1 & 1 \\ -1 & -1 & -1 \\ 0 & -2 & 0 \end{pmatrix} \Rightarrow A^{-1} = \begin{pmatrix} 0.5 & 0.5 & -0.5 \\ 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0 \end{pmatrix} .$$

Exercícios

7.3 -*Seja:*

$$A = \begin{pmatrix} 3 & 3 & -3 \\ -1 & 9 & 1 \\ 6 & 3 & -6 \end{pmatrix} .$$

Usando o método de Leverrier-Faddeev, determinar:

a) *seu polinômio característico,*

b) *seus auto-valores e correspondentes auto-vetores,*

c) A^{-1}.

7.4 - *Seja* $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, *definido por:*

$$T(x, y) = (3x + 5y, \ 3y) .$$

Usando o método de Leverrier-Faddeev, determinar seus auto-valores e correspondentes auto-vetores.

7.4 Método das Potências

O **Método das Potências** consiste em determinar o auto-valor de maior valor absoluto de uma matriz A, e seu correspondente auto-vetor, sem determinar o polinômio característico. O método é útil na prática, desde que se tenha interesse em determinar apenas alguns auto-valores, de módulo grande, e, que estes estejam bem separados, em módulo, dos demais. Podem surgir complicações caso a matriz A não possua auto-vetores linearmente independentes. O método das potências baseia-se no seguinte teorema.

Teorema 7.2 - *Seja A uma matriz real de ordem n e sejam $\lambda_1, \lambda_2, \ldots, \lambda_n$ seus auto-valores e $u_1, u_2, \ldots, u_n$ seus correspondentes auto-vetores. Suponha que os auto-vetores são linearmente independentes, e que:*

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| .$$

Seja a sequência y_k definida por:

$$y_{k+1} = A y_k , \quad k = 0, 1, 2, \ldots ,$$

onde y_0 é um vetor arbitrário, que permite a expansão:

$$y_0 = \sum_{j=1}^{n} c_j u_j ,$$

com c_j escalares quaisquer e $c_1 \neq 0$, então:

$$\lim_{k \to \infty} \frac{(y_{k+1})_r}{(y_k)_r} = \lambda_1 ,$$

onde o índice r indica a r-ésima componente. Além disso, quando $k \to \infty$, y_k tende ao auto-vetor correspondente a λ_1.

Prova: Temos por hipótese que:

$$y_0 = c_1 u_1 + c_2 u_2 + \ldots + c_n u_n \ . \tag{7.8}$$

Agora, lembrando que $Au_i = \lambda_i u_i$, obtemos:

$$\begin{aligned}
y_1 &= Ay_0 \\
&= c_1 A u_1 + c_2 A u_2 + \ldots + c_n A u_n \\
&= c_1 \lambda_1 u_1 + c_2 \lambda_2 u_2 + \ldots + c_n \lambda_n u_n \\
&= \lambda_1 \left[c_1 u_1 + c_2 \frac{\lambda_2}{\lambda_1} u_2 + \ldots + c_n \frac{\lambda_n}{\lambda_1} u_n \right] \ , \\
y_2 &= Ay_1 = A^2 y_0 \\
&= \lambda_1 \left[c_1 A u_1 + c_2 \frac{\lambda_2}{\lambda_1} A u_2 + \ldots + c_n \frac{\lambda_n}{\lambda_1} A u_n \right] \\
&= \lambda_1 \left[c_1 \lambda_1 u_1 + c_2 \frac{\lambda_2}{\lambda_1} \lambda_2 u_2 + \ldots + c_n \frac{\lambda_n}{\lambda_1} \lambda_n u_n \right] \\
&= \lambda_1^2 \left[c_1 u_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^2 u_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^2 u_n \right] \ , \\
&\vdots \\
y_k &= Ay_{k-1} = A^k y_0 \\
&= \lambda_1^k \left[c_1 u_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k u_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^k u_n \right] \ .
\end{aligned}$$

Desde que, por hipótese, $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$, temos então para $i = 1, \ldots, n$ que $\left| \frac{\lambda_i}{\lambda_1} \right| < 1$, e portanto quando $k \to \infty$, $\left(\frac{\lambda_i}{\lambda_1} \right)^k \to 0$.

Logo, o vetor:

$$\left[c_1 u_1 + c_2 \left(\frac{\lambda_2}{\lambda_1} \right)^p u_2 + \ldots + c_n \left(\frac{\lambda_n}{\lambda_1} \right)^p u_n \right] \ ,$$

converge para $c_1 u_1$ que é um múltiplo do auto-vetor correspondente ao auto-valor λ_1.

Assim, λ_1 é obtido de:

$$\lambda_1 = \lim_{k \to \infty} \frac{(y_{k+1})_r}{(y_k)_r} = \lim_{k \to \infty} \frac{\left(A^{k+1} y_0 \right)_r}{\left(A^k y_0 \right)_r} \quad , r = 1, 2, \ldots n \ . \tag{7.9}$$

e isso conclui a prova.

Observe então que, teoricamente, a partir de (7.9) obtemos o auto-valor de maior valor absoluto de uma matriz A. Na prática, para obter λ_1, utilizamos o algoritmo dado a seguir.

A partir de um vetor y_k, arbitrário, não nulo, construímos dois outros vetores y_{k+1} e z_{k+1}, do seguinte modo:

$$\begin{aligned}
z_{k+1} &= Ay_k \\
y_{k+1} &= \frac{1}{\alpha_{k+1}} z_{k+1}, \quad \text{onde } \alpha_{k+1} = \max_{1 \leq r \leq n} |\left(z_{k+1} \right)_r| \ ,
\end{aligned}$$

ou seja: dado um vetor y_0 qualquer, não nulo, construímos a sequência:

$$\begin{aligned}
z_1 &= Ay_0 \\
y_1 &= \frac{1}{\alpha_1} z_1 = \frac{1}{\alpha_1} A y_0 \\
z_2 &= Ay_1 = \frac{1}{\alpha_1} A^2 y_0 \\
y_2 &= \frac{1}{\alpha_2} z_2 = \frac{1}{\alpha_1 \alpha_2} A^2 y_0 \\
z_3 &= Ay_2 = \frac{1}{\alpha_1 \alpha_2} A^3 y_0 \\
&\vdots \\
y_k &= \frac{1}{\alpha_k} z_k = \frac{1}{\alpha_1 \alpha_2 \ldots \alpha_k} A^k y_0 \\
z_{k+1} &= Ay_k = \frac{1}{\alpha_1 \alpha_2 \ldots \alpha_k} A^{k+1} y_0 .
\end{aligned}$$

Assim, para obtermos λ_1, calculamos:

$$\lim_{k \to \infty} \frac{(z_{k+1})_r}{(y_k)_r} = \lim_{k \to \infty} \frac{\left(A^{k+1} y_0\right)_r}{(A^k y_0)_r} = \lambda_1 \ .$$

Observe que podemos garantir que o valor resultante fornece λ_1 desde que obtemos a mesma expressão dada por (7.9). Assim, pelo algoritmo, temos que:

$$\lim_{k \to \infty} \frac{(z_{k+1})_r}{(y_k)_r} = \lambda_1 \ . \tag{7.10}$$

Observações:

a) No limite, todas as componentes de $\frac{(z_{k+1})_r}{(y_k)_r}$ de (7.10), tendem a λ_1. Entretanto, na prática, uma das componentes converge mais rapidamente do que as outras. Assim, quando uma das componentes satisfizer a precisão desejada teremos o auto-valor procurado. Além disso, a velocidade de convergência depende de $\frac{\lambda_2}{\lambda_1}$. Portanto, quanto maior for $|\lambda_1|$ quando comparado com $|\lambda_2|$, mais rápida será a convergência.

b) Para obtermos λ_1 com uma precisão ϵ, em cada passo calculamos aproximações para λ_1 usando (7.10). O teste do erro relativo para cada componente de λ_1, isto é:

$$\frac{|\lambda_1^{(k+1)} - \lambda_1^{(k)}|_r}{|\lambda_1^{(k+1)}|_r} < \epsilon \ ,$$

é usado como critério de parada.

c) Quando todas as componentes de (7.10) forem iguais, então o vetor y_k dessa iteração é o auto-vetor correspondente ao auto-valor λ_1.

d) Se algum vetor resultar no vetor nulo, o método falha. Tal acontecimento deve ocorrer se as hipóteses não foram satisfeitas.

e) No Teorema 7.2 é feita a hipótese de $c_1 \neq 0$. Se $c_1 = 0$, então a prova do Teorema 7.2 indica que, teoricamente, o vetor y_k converge para u_2. Entretanto, na prática, para matrizes de ordem $n \geq 3$, que satisfaçam as demais condições do citado teorema, o método funciona sempre, pois, mesmo que o vetor y_0 não tenha componentes na direção de u_1, e desde que o método envolve a cada iteração uma divisão, os erros de arredondamento da máquina farão com que y_1 passe a ter componente nessa direção, após uma ou duas iterações.

Exemplo 7.4 - *Usando o método das potências determinar o auto-valor de maior valor absoluto da matriz:*

$$A = \begin{pmatrix} 3 & 0 & 1 \\ 2 & 2 & 2 \\ 4 & 2 & 5 \end{pmatrix},$$

com precisão de 10^{-2}.

Solução: Tomemos $y_0 = (1,\ 1,\ 1)^t$. Temos:

$$z_1 = Ay_0 = \begin{pmatrix} 4 \\ 6 \\ 11 \end{pmatrix} \ ; \ \alpha_1 = \max |\,(z_1)_r\,| = \max(|4|, |6|, |11|) = 11 \ .$$

$$y_1 = \frac{1}{\alpha_1} z_1 = \begin{pmatrix} 0.3636 \\ 0.5455 \\ 1 \end{pmatrix} \ , \ z_2 = Ay_1 = \begin{pmatrix} 2.0908 \\ 3.8182 \\ 7.5454 \end{pmatrix} .$$

Podemos então calcular uma 1ª aproximação para λ_1, usando (7.10). Logo:

$$\lambda_1^{(1)} = \frac{(z_2)_r}{(y_1)_r} = \begin{pmatrix} 5.7503 \\ 6.9995 \\ 7.5454 \end{pmatrix} .$$

Agora desde que $\alpha_2 = \max\{|2.0908|, |3.8182|, |7.5454|\} = 7.5454$, obtemos:

$$y_2 = \frac{1}{\alpha_2} z_2 = \begin{pmatrix} 0.2771 \\ 0.5060 \\ 1 \end{pmatrix} \ , \ z_3 = Ay_2 = \begin{pmatrix} 1.8313 \\ 3.5662 \\ 7.1204 \end{pmatrix} ,$$

Novamente, obtemos uma nova aproximação para λ_1, fazendo:

$$\lambda_1^{(2)} = \frac{(z_3)_r}{(y_2)_r} = \begin{pmatrix} 6.6088 \\ 7.0478 \\ 7.1204 \end{pmatrix} .$$

Calculando então o erro relativo, obtemos:

$$\frac{|\lambda_1^{(2)} - \lambda_1^{(1)}|_r}{|\lambda_1^{(2)}|_r} \simeq \begin{pmatrix} 0.13 \\ 0.07 \\ 0.13 \end{pmatrix} ,$$

o qual possui todas as componentes maiores que 10^{-2}. Assim, devemos fazer uma nova iteração. Agora desde que $\alpha_3 = 7.1204$, segue que:

$$y_3 = \frac{1}{\alpha_3} z_3 = \begin{pmatrix} 0.2572 \\ 0.5008 \\ 1 \end{pmatrix} \ , \ z_4 = Ay_3 = \begin{pmatrix} 1.8256 \\ 3.5160 \\ 7.0304 \end{pmatrix}$$

$$\Rightarrow \ \lambda_1^{(3)} = \frac{(z_4)_r}{(y_3)_r} = \begin{pmatrix} 7.0980 \\ 7.0208 \\ 7.0304 \end{pmatrix} .$$

Novamente, calculando o erro relativo:

$$\frac{|\lambda_1^{(3)} - \lambda_1^{(2)}|_r}{|\lambda_1^{(2)}|_r} \simeq \begin{pmatrix} 0.069 \\ 0.004 \\ 0.013 \end{pmatrix} ,$$

vemos que a segunda componente é menor que 10^{-2}. Portanto,

$$\lambda_1 \simeq 7.0208 \text{ com } \epsilon < 10^{-2} \text{ e } u_1 \simeq \begin{pmatrix} 0.2572 \\ 0.5008 \\ 1 \end{pmatrix} = y_3 .$$

Observações:

1) É claro que se desejamos λ_1 com precisão maior basta continuar fazendo iterações.

2) Os auto-valores de A são: $1, 2$ e 7 com auto-vetores: $(0.5,\ 1,\ -1)^t$, $(-1,\ 0.5,\ 1)^t$ e $(0.25,\ 0.5,\ 1)^t$, respectivamente.

3) O método das potências deve ser aplicado se o objetivo é determinar o auto-valor de maior valor absoluto de uma matriz. A desvantagem desse método é que ele fornece apenas um auto-valor de cada vez. Se todos os auto-valores são procurados devemos aplicar outros métodos que são muito mais eficientes.

4) Algumas vezes o maior auto-valor, em módulo, é o mais importante, mas se não é, devemos modificar o método. Em alguns problemas, o mais importante é a determinação do auto-valor de menor valor absoluto. Para isso dispomos da seguinte estratégia.

7.4.1 Método da Potência Inversa

O **Método da Potência Inversa** é usado para determinar o auto-valor de menor valor absoluto e seu correspondente auto-vetor de uma matriz A. O método é útil na prática, desde que se tenha interesse em determinar apenas o auto-valor, de menor módulo, e, que este esteja bem separado dos demais. Novamente, o método pode não funcionar caso a matriz A não possua auto-vetores linearmente independentes. O método da potência inversa é semelhante ao método das potências, com a diferença que agora assumimos:

$$|\lambda_1| \ \geq \ |\lambda_2| \ \geq \ \ldots | \ \geq \ \lambda_{n-1}| \ > \ |\lambda_n| \ ,$$

e desejamos determinar λ_n.

Sabemos que se λ é auto-valor de A, então λ^{-1} é auto-valor de A^{-1}. Além disso, se $|\lambda_n|$ é o menor auto-valor de A, então $|\lambda_n^{-1}|$ é o maior auto-valor de A^{-1}. Assim, o método da potência inversa consiste em calcular pelo método das potências o auto-valor de maior valor absoluto de A^{-1}, pois assim teremos o menor auto-valor, em módulo, de A. Portanto, dado y_k, construímos dois outros vetores y_{k+1} e z_{k+1} da seguinte forma :

$$\begin{aligned} z_{k+1} &= A^{-1} y_k \\ y_{k+1} &= \frac{1}{\alpha_{k+1}} z_{k+1}, \quad \text{onde } \alpha_{k+1} = \max_{1 \leq r \leq n} |\,(z_{k+1})_r\,| \ , \end{aligned}$$

e portanto:

$$\lambda_n^{-1} \ = \ \frac{(z_{k+1})_r}{(y_k)_r} \ .$$

Note que na prática não é necessário calcular A^{-1}, pois de:

$$z_{k+1} = A^{-1} y_k \Rightarrow A z_{k+1} = y_k \ ,$$

e assim resolvemos o sistema usando a **Decomposição LU**(ver Capítulo 4). Este método é particularmente conveniente desde que as matrizes L e U são independentes de k e portanto basta obtê-las uma única vez.

Exemplo 7.5 - *Deteminar o menor auto-valor, em módulo, da matriz:*

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 2 & 5 & 3 \\ 0 & 1 & 6 \end{pmatrix} ,$$

usando o método da potência inversa.

Solução: Os auto-valores de A são: $\lambda_1 = 7.44437$, $\lambda_2 = 4.21809$ e $\lambda_3 = 1.33754$. Portanto o maior auto-valor de A^{-1} é $\lambda_3^{-1} = \dfrac{1}{1.33754} \simeq 0.7476$, e é esse valor que desejamos encontrar.

Decompondo A em LU, obtemos:

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0.25 & 1 \end{pmatrix} , \quad U = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 4 & 3 \\ 0 & 0 & 5.25 \end{pmatrix} .$$

Assim, tomando $y_0 = (1,\ 1,\ 1)^t$ em $Az_1 = y_0$ ou seja fazendo$LUz_1 = y_0$, segue que:

$$z_1 = \begin{pmatrix} 0.5715 \\ -0.1429 \\ 0.1905 \end{pmatrix} , \ \alpha_1 = 0.5715 \ , \quad y_1 = \frac{1}{\alpha_1} z_1 = \begin{pmatrix} 1 \\ -0.2500 \\ 0.3333 \end{pmatrix} .$$

Resolvendo agora $LUz_2 = y_1$, obtemos:

$$z_2 = \begin{pmatrix} 0.7024 \\ -0.4048 \\ 0.1230 \end{pmatrix} \quad \Rightarrow \lambda_3^{-1} = \frac{(z_2)_r}{(y_1)_r} = \begin{pmatrix} 0.7024 \\ 1.6192 \\ 0.3690 \end{pmatrix} .$$

Agora, $\alpha_2 = 0.7024$. Continuando o processo, obtemos:

$$y_2 = \frac{1}{\alpha_2} z_2 = \begin{pmatrix} 1 \\ -0.5763 \\ 0.1751 \end{pmatrix} , \text{ e de } LUz_3 = y_2 \Rightarrow z_3 = \begin{pmatrix} 0.7377 \\ -0.4754 \\ 0.1084 \end{pmatrix}$$

$$\Rightarrow \lambda_3^{-1} = \frac{(z_3)_r}{(y_2)_r} = \begin{pmatrix} 0.7377 \\ 0.8249 \\ 0.6192 \end{pmatrix} . \text{ Temos} : \alpha_3 = 0.7377, \text{ e assim} :$$

$$y_3 = \frac{1}{\alpha_3} z_3 = \begin{pmatrix} 1 \\ -0.6444 \\ 0.1469 \end{pmatrix} \text{ e de } LUz_4 = y_3 \Rightarrow z_4 = \begin{pmatrix} 0.7454 \\ -0.4908 \\ 0.1063 \end{pmatrix} .$$

$$\Rightarrow \lambda_3^{-1} = \frac{(z_4)_r}{(y_3)_r} = \begin{pmatrix} 0.7454 \\ 0.7617 \\ 0.7235 \end{pmatrix} . \text{ Finalmente}, \alpha_4 = 0.7454, \text{ e portanto} :$$

$$y_4 = \frac{1}{\alpha_4} z_4 = \begin{pmatrix} 1 \\ -0.6584 \\ 0.1426 \end{pmatrix} \text{ e de } LUz_5 = y_4 \Rightarrow z_5 = \begin{pmatrix} 0.7471 \\ -0.4942 \\ 0.1061 \end{pmatrix},$$

$$\Rightarrow \lambda_3^{-1} = \frac{(z_5)_r}{(y_4)_r} = \begin{pmatrix} 0.7471 \\ 0.7506 \\ 0.7443 \end{pmatrix}.$$

Logo $\lambda_3^{-1} \simeq 0.7471$ é o auto-valor de maior valor absoluto de A^{-1}. Portanto $\frac{1}{\lambda_3^{-1}} \simeq 1.3385$ é o auto-valor de menor valor absoluto de A.

7.4.2 Método das Potências *com Deslocamento*

Suponha agora que A tem auto-valores λ_i, reais, com

$$\lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_{n-1} > \lambda_n .$$

e considere a sequência de vetores definida por:

$$z_{k+1} = (A - qI)y_k$$

$$y_{k+1} = \frac{1}{\alpha_{k+1}} z_{k+1}, \quad \text{onde } \alpha_{k+1} = \max_{1 \leq r \leq n} |(z_{k+1})_r|,$$

onde I é a matriz identidade de ordem n e q é um parâmetro qualquer. Isto é chamado **Método das Potências** *com Deslocamento*, porque $A - qI$ tem auto-valores $\lambda_i - q$, isto é, os auto-valores de A são *deslocados* q unidades na reta real. Os auto-vetores de $A - qI$ são os mesmos da matriz A.

Portanto o Teorema 7.2 pode ser aplicado à matriz $A - qI$, e pode ser mostrado que y_k converge para o auto-vetor correspondente àquele que maximiza $|\lambda_i - q|$. Portanto se:

$$q < \frac{\lambda_1 + \lambda_n}{2} \text{ então } y_k \to u_1 \text{ e } \lim_{k \to \infty} \frac{(z_{k+1})_r}{(y_k)_r} \to \lambda_1 - q,$$

$$q > \frac{\lambda_1 + \lambda_n}{2} \text{ então } y_k \to u_n \text{ e } \lim_{k \to \infty} \frac{(z_{k+1})_r}{(y_k)_r} \to \lambda_n - q,$$

Assim, a escolha apropriada de q pode ser usada para determinar os dois auto-valores extremos, correspondendo ao maior e ao menor auto-valor de A. Observe que se $q = (\lambda_1 + \lambda_n)/2$ então $\lambda_1 - q = -(\lambda_n - q)$, e assim $A - qI$ tem dois auto-valores de mesmo módulo, mas de sinais opostos. Neste caso, a sequência de vetores oscilará entre dois limites os quais são duas combinações de u_1 e u_2.

O auto-valor e o auto-vetor *dominante* são usualmente calculados tomando um *deslocamento* **zero**, isto é, o cálculo para determinar λ_1 e u_1 são realizados na matriz A, através do método das potências. A matriz pode então ser *deslocada* de λ_1 para estimar o auto-valor λ_n.

Exemplo 7.6 - *Determinar o auto-valor de menor valor absoluto da matriz dada no exemplo 7.4, usando o método das potências com deslocamento.*

Solução: No exemplo 7.4, o auto-valor de maior valor absoluto foi estimado $\simeq 7$. Assim, para determinar o auto-valor de menor valor absoluto, vamos aplicar o método das potências na matriz:

$$A - 7I = \begin{pmatrix} -4 & 0 & 1 \\ 2 & -5 & 2 \\ 4 & 2 & -2 \end{pmatrix} = A^*.$$

Iniciando com $y_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, obtemos:

$$z_1 = A^* y_0 = \begin{pmatrix} -3 \\ -1 \\ 4 \end{pmatrix} \; ; \; \alpha_1 = \max |(z_1)_r| = 4 \; .$$

$$y_1 = \frac{1}{\alpha_1} z_1 = \begin{pmatrix} -0.75 \\ -0.25 \\ 1 \end{pmatrix} \; , \; z_2 = A^* y_1 = \begin{pmatrix} 4.00 \\ 1.75 \\ -5.50 \end{pmatrix} \; .$$

Podemos, então, calcular uma primeira aproximação para λ_1^*. Assim:

$$\lambda_1^{*(1)} = \frac{(z_2)_r}{(y_1)_r} = \begin{pmatrix} -5.33 \\ -7.00 \\ -5.50 \end{pmatrix} \; .$$

Continuando o processo, obteremos:

$$y_{19} = \begin{pmatrix} -0.52 \\ -0.94 \\ 1 \end{pmatrix} \; , \; z_{20} = A^* y_{19} = \begin{pmatrix} 3.03 \\ 5.71 \\ -5.98 \end{pmatrix}$$

$$\Rightarrow \quad \lambda_1^{*(19)} = \frac{(z_{20})_r}{(y_{19})_r} = \begin{pmatrix} -5.92 \\ -5.95 \\ -5.98 \end{pmatrix} \; .$$

Assim, podemos concluir que o auto-valor *dominante* de A^* é aproximadamente -5.98 com auto-vetor aproximado $u_1^* = (-0.52, \; -0.94, \; 1)^t$. Portanto a matriz original possui o mesmo auto-vetor mas seu auto-valor é $-5.98 + 7.00 = 1.02$. A lentidão na convergência neste caso se deve ao fato que os auto-valores de A^* são: $-6, -5$ e 0 e assim a convergência é governada pelo fator: $\left(\frac{5}{6}\right)^k$. Compare com o exemplo 7.4, e 7.5, onde a razão de convergência é $\left(\frac{2}{7}\right)^k$ e $\left(\frac{1.33754}{4.21809}\right)^k$, respectivamente.

Em geral, se $y_k \to u_1$, então na presença do *deslocamento* q, a velocidade de convergência depende de:

$$\left(\frac{\lambda_i - q}{\lambda_1 - q}\right)^k \; ,$$

e assim uma escolha adequada de q pode acelerar a convergência. Por exemplo, se A é uma matriz de ordem 3, com auto-valores: $5, 7$ e 10, sem *deslocamento* a convergência depende de $\left(\frac{7}{10}\right)^k$, mas com um *deslocamento* de 6 dependerá de $\left(\frac{1}{4}\right)^k$, pois $A - 6I$ tem auto-valores: $-1, 1$ e 4.

Portanto, na prática não é trivial encontrar o melhor valor de q, a menos que alguns dos auto-valores sejam conhecidos a priori. O método das potências e /ou o método das potências *com deslocamento* devem ser utilizados se apenas um ou dois dos auto-valores são desejados. Se o objetivo é determinar mais auto-valores então o método da potência inversa *com deslocamento* pode ser usado, ou seja, como no método da potência inversa, calculamos:

$$(A - qI)z_{k+1} \; = \; y_k \; ,$$

usando a decomposição LU, e assim os auto valores de $(A-qI)^{-1}$ serão $\dfrac{1}{(\lambda_i - q)}$. Novamente, o Teorema 7.2 pode ser aplicado a $(A - qI)^{-1}$ e deduzimos que y_k converge para o auto-vetor correspondente ao auto-valor que maximiza $\dfrac{1}{|\lambda_i - q|}$. Escolhas adequadas dos valores de q nos permitem determinar todos os auto-valores de A, e não somente aqueles correspondentes aos auto-valores extremos. Assim, se o auto-valor próximo a q é λ_j, então o valor de λ_j pode ser calculado a partir de:

$$\bar{\lambda}_j \; = \; \frac{1}{(\lambda_j - q)} \; ,$$

onde $\bar{\lambda}_j$ é o auto-valor de $(A - qI)^{-1}$, obtido pelo método da potência inversa *com deslocamento* q.

Exemplo 7.7 - *Determinar o segundo maior auto-valor, em valor absoluto, da matriz dada no exemplo 7.4.*

Solução: Já determinamos dois auto-valores desta matriz: 7 e 1.02 (Exemplos 7.4 e 7.6). Sabemos que o **traço** de uma matriz é igual a soma dos seus auto-valores . Neste exemplo o traço de A é 10 e assim o outro auto-valor é aproximadamente 1.98, o qual será tomado como o valor de q na iteração inversa *com deslocamento*. Assim, montamos a matriz:

$$A - 1.98I \; = \; \begin{pmatrix} 1.02 & 0 & 1 \\ 2 & 0.02 & 2 \\ 4 & 2 & 3.02 \end{pmatrix} \; ,$$

e a decompomos no produto LU, onde:

$$L \; = \; \begin{pmatrix} 1 & & \\ 1.9608 & 1 & \\ 3.9216 & 100 & 1 \end{pmatrix} \; , \; U \; = \; \begin{pmatrix} 1.02 & 0 & 1 \\ & 0.02 & 0.0392 \\ & & -4.8216 \end{pmatrix} .$$

Tomando como vetor inicial $y_0 = (1, \; 1, \; 1)^t$, e resolvendo o sistema linear $LUz_1 = y_0$, resulta :

$$z_1 \; = \; \begin{pmatrix} 19.9226 \\ -10.1707 \\ -19.3211 \end{pmatrix} \Rightarrow y_1 = \frac{1}{19.9226} z_1 \; = \; \begin{pmatrix} 1 \\ -0.5105 \\ -0.9698 \end{pmatrix}$$

De $LUz_2 = y_1$, obtemos:

$$z_2 \; = \; \begin{pmatrix} 50.2356 \\ -25.0940 \\ -50.2403 \end{pmatrix} \Rightarrow \lambda_2^{*(1)} \; = \frac{z_2}{y_1} \; = \; \begin{pmatrix} 50.2356 \\ 49.0500 \\ 51.8048 \end{pmatrix} .$$

Agora,

$$y_2 = \frac{1}{-50.2403} z_2 \; = \; \begin{pmatrix} -0.9999 \\ 0.4995 \\ 1 \end{pmatrix} .$$

Fazendo $LUz_3 = y_2$, obtemos:

$$z_3 = \begin{pmatrix} -50.4088 \\ 24.1885 \\ 50.4166 \end{pmatrix} \rightarrow \lambda_2^{*(2)} = \frac{z_3}{y_2} = \begin{pmatrix} 50.4138 \\ 48.3180 \\ 51.4166 \end{pmatrix} .$$

Assim, $\lambda_2^* \simeq 50.41$. Portanto, o segundo maior auto-valor, em valor absoluto de A é:

$$\lambda_2 = 1.98 + \frac{1}{50.41} = 1.9998 .$$

Observe que o sucesso do método das potências *com deslocamento* depende de nossa habilidade em obter estimativas precisas para usar no *deslocamento*. Neste último exemplo, uma estimativa para λ_2 foi obtida usando a relação entre o traço da matriz e a soma dos auto-valores. Infelizmente, para matrizes de ordem > 3, não é fácil obter valores apropriados para os *deslocamentos*. Como já dissemos anteriormente, se desejamos todos os auto-valores devemos usar outros métodos.

Exercícios

7.5 - *Determinar o auto-valor de maior valor absoluto e seu correspondente auto-vetor, da matriz:*

$$A = \begin{pmatrix} 1 & -1 & 3 \\ -1 & 1 & 3 \\ 3 & -3 & 9 \end{pmatrix} .$$

, calculando apenas a primeira aproximação pelo método das potências. O qua você pode concluir?

7.6 - *Usando o método das potências calcular, o auto-valor de maior valor absoluto e seu correspondente auto-vetor, da matriz:*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} .$$

com precisão de 10^{-2}.

7.7 - *Usando o método da potência inversa, calcule o auto-valor de menor valor absoluto da matriz:*

$$A = \begin{pmatrix} 2 & 4 & -2 \\ 4 & 2 & 2 \\ -2 & 2 & 5 \end{pmatrix} ,$$

com precisão de 10^{-2}.

7.8 - *Sabendo que o auto-valor de maior valor absoluto da matriz:*

$$A = \begin{pmatrix} 4 & -1 & 1 \\ 1 & 1 & 1 \\ -2 & 0 & -6 \end{pmatrix} ,$$

é aproximadamente: -5.76849*, e que seu correspondente auto-vetor é aproximadamente:* $(-0.1157,\ -0.1306,\ 1)^t$*, calcule os demais auto-valores e correspondentes auto-vetores de* A*, usando:*

a) *o método das potências com deslocamento para obter o menor auto-valor, em valor absoluto,*

b) *o método da potência inversa com deslocamento para obter o auto-valor* λ_2.

7.5 Auto-Valores de Matrizes Simétricas

Nessa seção restringiremos nossa atenção para matrizes simétricas de ordem n. Matrizes deste tipo possuem auto-valores reais e os auto-vetores são linearmente independentes. O método de Jacobi, que descreveremos mais adiante, é usado para determinar os auto-valores e auto-vetores, de matrizes simétricas, através de uma série de transformações similares:

$$A_{k+1} \; = \; U_k^{-1} A_k U_k \; , \quad k = 1, 2, \ldots \; ,$$

onde $A_1 = A$. As matrizes $A_1, A_2, \ldots$ convergem num número infinito de passos para uma matriz diagonal. Os auto-valores e auto-vetores são então determinados em virtude do Lema 1.1 (o qual se aplica tanto para matrizes simétricas como para matrizes não simétricas).

Assim, após m passos do método de Jacobi, obteremos:

$$A_{m+1} = U_m^{-1} \ldots U_2^{-1} U_1^{-1} A_1 U_1 U_2 \ldots U_m \; .$$

Portanto, se $A_{m+1} \simeq D$, segue que os elementos diagonais de A_{m+1} são aproximações para os auto-valores de A e as colunas de $V = U_1 U_2 \ldots U_m$ são aproximações para os auto-vetores.

Para descrevermos o método de Jacobi, (para matrizes simétricas), precisamos de alguns conceitos, os quais passamos a considerar agora. Assim:

Rotação de Jacobi

Seja A uma matriz simétrica. Uma rotação (p, q) de Jacobi é a operação $U^t A U$ com U dada por (1.23). Observe que fazer uma rotação de Jacobi é efetuar uma transformação de semelhança na matriz A.

Para um melhor entendimento, consideremos inicialmente, uma rotação (2,4) de Jacobi, em uma matriz A de ordem 4. Efetuando o produto $U^t A$, obtemos:

$$U^t A \; = \; \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\;\varphi & 0 & -sen\;\varphi \\ 0 & 0 & 1 & 0 \\ 0 & sen\;\varphi & 0 & cos\;\varphi \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21}c - a_{41}s & a_{22}c - a_{42}s & a_{23}c - a_{43}s & a_{24}c - a_{44}s \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{21}s - a_{41}c & a_{22}s + a_{42}c & a_{23}s - a_{43}c & a_{24}s - a_{44}c \end{pmatrix}$$

$$= \; A' \; = \; (a'_{ij}), \text{ onde } cos\;\varphi = c \text{ e } \; sen\;\varphi = s.$$

Fazendo agora o produto $A'U$, segue que:

$$A'U \; = \; \begin{pmatrix} a'_{11} & a'_{12} & a'_{13} & a'_{14} \\ a'_{21} & a'_{22} & a'_{23} & a'_{24} \\ a'_{31} & a'_{32} & a'_{33} & a'_{34} \\ a'_{41} & a'_{42} & a'_{43} & a'_{44} \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & \\ 0 & cos\;\varphi & 0 & sen\;\varphi \\ 0 & 0 & 1 & 0 \\ 0 & -sen\;\varphi & 0 & cos\;\varphi \end{pmatrix}$$

$$= \begin{pmatrix} a'_{11} & a'_{12}c - a'_{14}s & a'_{13} & a'_{12}s + a'_{14}c \\ a'_{21} & a'_{22}c - a'_{24}s & a'_{23} & a'_{22}s + a'_{24}c \\ a'_{31} & a'_{32}c - a'_{34}s & a'_{33} & a'_{32}s + a'_{34}c \\ a'_{41} & a'_{42}c - a'_{44}s & a'_{43} & a'_{42}s + a'_{44}c \end{pmatrix} \; = \; A'' \; = \; (a''_{ij}) \; .$$

Assim, de um modo geral, para uma matriz de ordem n o produto $U^t A$, fornece uma matriz A', onde:

$$\begin{cases} a'_{pj} \ = \ a_{pj} \ cos \ \varphi \ - \ a_{qj} \ sen \ \varphi \ , & 1 \leq j \leq n \ , \\ a'_{qj} \ = \ a_{pj} \ sen \ \varphi \ + \ a_{qj} \ cos \ \varphi \ , & 1 \leq j \leq n \ , \\ a'_{ij} = a_{ij} \ , \qquad i \neq p, q \ , & 1 \leq j \leq n \ . \end{cases} \tag{7.11}$$

e o produto $A'U$ fornece uma matriz A'', onde:

$$\begin{cases} a''_{ip} \ = \ a'_{ip} \ cos \ \varphi \ - \ a'_{iq} \ sen \ \varphi \ , & i \leq i \leq n \ , \\ a''_{iq} \ = \ a'_{ip} \ sen \ \varphi \ + \ a'_{iq} \ cos \ \varphi \ , & i \leq i \leq n \ , \\ a''_{ij} \ = \ a'_{ij} \ , \quad j \neq p, q \ , & i \leq i \leq n \ . \end{cases} \tag{7.12}$$

Portanto, a matriz A'' tem a seguinte forma:

$$A'' = \begin{pmatrix} \ddots & \vdots & & \vdots & & \\ \ldots & \bigcirc & \ldots & \bigcirc & \ldots & p \\ & \vdots & \ddots & \vdots & & \\ \ldots & \bigcirc & \ldots & \bigcirc & \ldots & q \\ & \vdots & & \vdots & \ddots & \\ & p & & q & & \end{pmatrix} ,$$

isto é, na matriz A'' apenas os elementos das linhas e colunas p e q serão alterados, sendo que os elementos $a_{pp}, \ a_{pq}, \ a_{qp}, \ a_{qq}$ serão transformados duas vezes. Portanto A'' continua simétrica.

Vejamos agora as fórmulas que determinam a passagem de $A \rightarrow A''$, denominada **Rotação de Jacobi** de um ângulo φ para os elementos da interseção. Temos, utilizando (7.12) e (7.11), que:

$$\begin{aligned} 1) \ a''_{pp} \ &= \ a'_{pp} \ cos \ \varphi \ - \ a'_{pq} \ sen \ \varphi \\ &= \ (a_{pp} \ cos \ \varphi \ - \ a_{qp} \ sen \ \varphi) \ cos \ \varphi \ - \\ &- \ (a_{pq} \ cos \ \varphi \ - \ a_{qq} \ sen \ \varphi) \ sen \ \varphi \ . \end{aligned}$$

Portanto:

$$a''_{pp} = a_{pp} \ cos^2 \ \varphi \ - \ 2a_{pq} \ sen \ \varphi \ cos \ \varphi \ + \ a_{qq} \ sen^2 \ \varphi \ . \tag{7.13}$$

$$\begin{aligned} 2) \ a''_{qq} \ &= \ a'_{gp} \ sen \ \varphi \ + \ a'_{qq} \ cos \ \varphi \\ &= \ (a_{pp} \ sen \ \varphi \ + \ a_{qp} \ cos \ \varphi) \ sen \ \varphi \ + \\ &+ \ (a_{pq} \ sen \ \varphi \ - \ a_{qq} \ cos \ \varphi) \ cos \ \varphi \ . \end{aligned}$$

Logo:

$$a''_{qq} = a_{pp} \ sen^2 \ \varphi \ + \ 2a_{pq} \ sen \ \varphi \ cos \ \varphi \ + \ a_{qq} \ cos^2 \ \varphi \ . \tag{7.14}$$

$$\begin{aligned} 3) \ a''_{pq} \ &= \ a'_{pp} \ sen \ \varphi \ + \ a'_{pq} \ cos \ \varphi \\ &= \ (a_{pp} \ cos \ \varphi \ - \ a_{qp} \ sen \ \varphi) \ sen \ \varphi \ + \\ &+ \ (a_{pq} \ cos \ \varphi \ - \ a_{qq} \ sen \ \varphi) \ cos \ \varphi \ . \end{aligned}$$

Assim:

$$a''_{pq} = a''_{qp} \ = \ (a_{pp} - a_{qq}) sen \ \varphi \ cos \ \varphi \ + \ a_{pq} \ (cos^2 \ \varphi - sen^2 \ \varphi) \ . \tag{7.15}$$

Portanto, para fazer uma rotação (p, q)de Jacobi, usamos as fórmulas: (7.13), (7.14), (7.15), (7.12) com $j \neq p, q$ e (7.11) com $i \neq p, q$.

Exemplo 7.8 - *Considere a matriz:*

$$A = \begin{pmatrix} 2 & 1 & 3 & 1 \\ 1 & 0 & -1 & 0 \\ 3 & -1 & 3 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} .$$

Fazer uma rotação de $\varphi = \frac{\pi}{2}$ *em torno do elemento* $(p,q) = (1,3)$.

Solução: Temos:

$$\begin{aligned} cos\ \varphi &= cos\ 90^\circ = 0 , \\ sen\ \varphi &= sen\ 90^\circ = 1 . \end{aligned}$$

Agora, utilizando as fórmulas anteriores, obtemos:

$$\begin{aligned} \text{de } (7.7) \Rightarrow a''_{11} &= a_{11}\ c^2 - 2a_{13}\ s\ c + a_{33}\ s^2 = a_{33} = 3 , \\ \text{de } (7.8) \Rightarrow a''_{33} &= a_{11}\ s^2 + 2a_{13}\ s\ c + a_{33}\ c^2 = a_{11} = 2 , \\ \text{de } (7.9) \Rightarrow a''_{13} &= a''_{31} = (a_{11} - a_{33})\ s\ c + a_{13}\ (c^2 - s^2) = -a_{13} = -3 . \end{aligned}$$

Usando (7.12) e (7.11), segue que:

$$\begin{aligned} a''_{12} &= a'_{12} = a_{12}\ c - a_{32}\ s = -a_{32} = 1 = a''_{21} , \\ a''_{14} &= a'_{14} = a_{14}\ c - a_{34}\ s = -a_{34} = 0 = a''_{41} , \\ a''_{32} &= a'_{32} = a_{12}\ s + a_{32}\ c = a_{12} = 1 = a''_{23} , \\ a''_{34} &= a'_{34} = a_{14}\ s + a_{34}\ c = a_{14} = 1 = a''_{43} . \end{aligned}$$

Assim:

$$A'' = \begin{pmatrix} 3 & 1 & -3 & 0 \\ 1 & 0 & 1 & 0 \\ -3 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} ,$$

corresponde a uma rotação de 90° em torno do elemento (1,3).

7.5.1 Método Clássico de Jacobi

O **Método Clássico de Jacobi**, ou simplesmente **Método de Jacobi**, como já dissemos, é um método numérico que serve para determinar auto-valores e auto-vetores de matrizes simétricas. Dada a matriz A, efetuamos uma sequência de rotações:

$$A_1 = A\ ;\ A_2 = U_1^t\ A_1\ U_1\ \rightarrow\ A_3 = U_2^t\ A_2\ U_2\ \rightarrow$$

$$\rightarrow \ldots\ \rightarrow\ A_{k+1} = U_k^t\ A_k\ U_k \simeq D ,$$

onde $U_i,\ i = 1, 2 \ldots k$ são matrizes de rotação, e D é uma matriz diagonal.

O processo para construção da matriz A_2, consiste em escolhermos entre os elementos não diagonais de A o elemento de maior valor absoluto, isto é:

$$a_{pq} = \max_{i \neq j}(a_{ij}) .$$

Fazer então, uma rotação com a finalidade de zerar o elemento a_{pq}. A seguir reaplicamos o processo à matriz resultante tantas vezes quantas forem necessárias, de tal modo a reduzirmos a matriz A a uma matriz diagonal D, cujos elementos são os auto-valores de A.

Assim, no primeiro passo devemos zerar o elemento a_{pq}. Assumimos que $a_{pq} \neq 0$, (pois caso contrário nada teríamos a fazer), e assim nosso objetivo é obter $a''_{pq} = 0$. De (7.15), temos a expressão para a''_{pq} e impondo que o mesmo seja identicamente nulo, segue que:

$$(a_{pp} - a_{qq}) \underbrace{cos\ \varphi\ sen\ \varphi}_{\frac{1}{2} sen 2\varphi} + apq(\underbrace{cos^2\ \varphi - sen^2\ \varphi}_{cos 2\varphi}) = 0\ .$$

Portanto:

$$\begin{aligned} a_{pp}\ -\ a_{qq} &= -\ \frac{a_{pq}\ cos\ 2\ \varphi}{\frac{1}{2}\ sen\ 2\ \varphi} = -\ 2\ a_{pq}\ cotg\ 2\ \varphi \\ \Rightarrow\ cotg\ 2\ \varphi &= \frac{aqq\ -\ a_{pp}}{2\ a_{pq}} = \phi\ . \end{aligned}$$

Agora:

$$\begin{aligned} cotg\ 2\ \varphi &= \frac{cos\ 2\ \varphi}{sen\ 2\ \varphi} = \frac{cos^2\ \varphi\ -\ sen^2\ \varphi}{2\ sen\ \varphi\ cos\ \varphi} = \\ &= \frac{\frac{cos^2\ \varphi\ -\ sen^2\ \varphi}{cos^2\ \varphi}}{\frac{2\ sen\ \varphi\ cos\ \varphi}{cos^2\ \varphi}} = \frac{1\ -\ tg^2\ \varphi}{2\ tg\ \varphi}\ . \end{aligned}$$

Seja $t = tg\ \varphi$; temos $cotg\ 2\ \varphi\ =\ \phi$. Assim:

$$\phi\ =\ \frac{1 - t^2}{2t}\ \Rightarrow\ 1 - t^2\ =\ 2t\phi\ .$$

Portanto:

$$t^2\ +\ 2t\phi - 1 = 0\ \Rightarrow \Rightarrow\ t\ =\ \frac{-2\ \phi \pm\ \sqrt{4\phi^2 + 4}}{2}\ .$$

Obtemos então: $t = -\phi \pm \sqrt{\phi^2 + 1}$. Multiplicando o numerador e o denominador por: $\phi \pm \sqrt{\phi^2 + 1}$ segue que:

$$t\ =\ \frac{1}{\phi \pm \sqrt{\phi^2 + 1}}$$

Computacionalmente, adotamos:

$$t\ =\ \begin{cases} \dfrac{1}{\phi + Sinal(\phi)\sqrt{\phi^2 + 1}}\ , & \phi \neq 0\ ; \\ \\ 1\ , & \phi = 0\ . \end{cases}$$

Observe que escolhemos o sinal positivo ou negativo de ϕ de modo a obter o denominador de maior módulo, pois assim teremos sempre $|t| \leq 1$. Agora, temos as seguintes fórmulas para a secante de um ângulo φ:

$$sec^2\ \varphi\ =\ 1 + tg^2\ \varphi, \quad \text{e}\ , \quad sec^2\ \varphi = \frac{1}{cos^2\ \varphi}\ .$$

Assim:

$$\frac{1}{cos^2\ \varphi}\ =\ 1 + tg^2\ \varphi\ \Rightarrow\ cos^2\ \varphi\ =\ \frac{1}{1 + tg^2\ \varphi}\ .$$

Logo, podemos escrever:

$$\begin{aligned} c &= cos\ \varphi = \frac{1}{\sqrt{1+tg^2\ \varphi}} = \frac{1}{\sqrt{1+t^2}}\ , \\ s &= sen\ \varphi = cos\ \varphi \cdot t = \frac{t}{\sqrt{1+t^2}}\ . \end{aligned}$$

Resumindo, o método de Jacobi, consiste em:

1) Determinar o elemento de maior módulo de A fora da diagonal. Esse elemento será denotado por a_{pq}.

2) Calcular:

2.1) $\phi = \dfrac{a_{qq} - a_{pp}}{2a_{pq}}$.

2.2) $t = \begin{cases} \dfrac{1}{\phi + Sinal(\phi)\sqrt{\phi^2+1}}\ , & \phi \neq 0\ ; \\ 1\ , & \phi = 0\ . \end{cases}$

2.3) $cos\ \varphi = \dfrac{1}{\sqrt{1+t^2}}$.

2.4) $sen\ \varphi = \dfrac{t}{\sqrt{1+t^2}}$.

3) Usar as fórmulas de rotação de Jacobi, isto é: as fórmulas: (7.13), (7.14), (7.6) com $j \neq p, q$ e (7.5) com $i \neq p, q$.

O processo deve ser repetido até obtermos uma matriz diagonal.

Observe que em cada passo k, o item **3)** acima pode ser substituído pelo produto $U_k^t A_k U_k$.

Cálculo dos Auto-Vetores

Ao mesmo tempo que calculamos os auto-valores de uma matriz A pelo método de Jacobi podemos obter seus auto-vetores. Vimos que a sequência de matrizes A_k é calculada por recorrência através de:

$$A_{k+1} = U_k^t\ A_k U_k \quad (k = 1, 2, \ldots).$$

Como $A_1 = A$, obtemos:

$$A_{k+1} = U_k^t\ U_{k-1}^t \ldots U_2^t U_1^t\ A\ U_1 U_2 \ldots U_{k-1} U_k = V^t A V,$$

onde $V = U_1 U_2 \ldots U_{k-1} U_k$.

Com a hipótese que $A_k \simeq D$ obtemos que $D = V^t A V$, onde V é matriz ortogonal, pois a matriz V é produto de matrizes ortogonais. Assim D contém os auto-valores de A e V contém seus correspondentes auto-vetores (em colunas), isto é, a j-ésima coluna de V é o auto-vetor correspondente ao auto-valor λ_j.

Observe que em cada passo do método de Jacobi, um par de elementos fora da diagonal torna-se zero. Assim pode parecer, à primeira vista, que uma matriz diagonal é obtida após um número finito de passos. Entretanto, isso não é verdade porque transformações ortogonais subsequentes destroem os zeros criados

anteriormente. Apesar disso, é possível mostrar que quando um zero é criado nas posições (p,q) e (q,p), a soma dos quadrados dos elementos não diagonais da matriz A_k, $S(A_k)$, decresce de $2a_{pq}^2$. De fato, seja:

$$S(A_k) \;=\; \sum_{\substack{i,j=1\\ i\neq j}} (a_{ij})^2 \; .$$

Vamos mostrar que $S(A_k) \to 0$. Para tanto, em cada passo $A \to A''$ vamos comparar $S(A)$ com $S(A'')$. Assim:

$$\begin{aligned} S(A'') \;&=\; \sum_{\substack{i,j=1\\ i,j\neq p,q}} (a''_{ij})^2 + \sum_{\substack{i=1\\ i\neq p,q}} [(a''_{ip})^2 \\ &+\; (a''_{iq})^2] + \sum_{\substack{j=1\\ j\neq p,q}} [(a''_{pj})^2 + (a''_{qj})^2] + 2(a''pq)^2 \; , \end{aligned}$$

onde as somas do lado direito da expressão acima representam, respectivamente: os elementos que não mudam, os elementos das linhas p e q, fora da diagonal; elementos das colunas p e q, fora da diagonal. Agora, usando (**??**), segue que:

$$(a''_{ip})^2 + (a''_{iq})^2 \;=\; (a_{ip}c - a_{iq}s)^2 + (a_{ip}s - a_{iq}c)^2 \;=\; (a_{ip})^2 + (a_{iq})^2 \; ,$$

e desde que o mesmo é válido para $(a''_{pj})^2 + (a''_{qj})^2$, obtemos:

$$S(A'') \;=\; S(A) - 2(a_{pq})^2 + 2(a''_{pq})^2 \; .$$

Observe que na expressão acima devemos subtrair $2(a_{pq})^2$, pois $S(A)$ contém este elemento. Assim, de um modo geral, no k-ésimo passo, teremos:

$$\begin{aligned} S_k \;&=\; S_{k-1} - 2(a_{pq}^{k-1})^2 + 2(a_{pq}^{k})^2 \\ &=\; S_{k-1} - 2(a_{pq}^{k-1})^2 \; , \end{aligned}$$

desde que (a_{pq}^{k-1}) é o maior elemento, em módulo, fora da diagonal principal e $a_{pq}^k = 0$. Substituindo todos os elementos, fora da diagonal principal, por $a_{pq}^{k-1})$, obtemos:

$$\begin{aligned} S_{k-1} \;&\leq\; (n^2-n)(a_{pq}^{k-1})^2 \\ \Rightarrow (a_{pq}^{k-1})^2 \;&\leq\; \frac{S_{k-1}}{n^2-n} \; . \end{aligned}$$

Logo:

$$\begin{aligned} S_k \;&\leq\; S_{k-1} - 2\frac{S_{k-1}}{n^2-n} \\ &=\; S_{k-1}\left(1 - \frac{2}{n^2-n}\right) \; . \end{aligned}$$

A partir desta expressão para S_k, podemos escrever que:

$$S_k \leq \left(1 - \frac{2}{n^2-n}\right) S_{k-1} \leq \left(1 - \frac{2}{n^2-n}\right)^2 S_{k-2} \leq \ldots \; ,$$

e assim, concluímos que:

$$S_k \leq \left(1 - \frac{2}{n^2 - n}\right)^k S_0 \ ,$$

onde S_0, representa a soma dos quadrados dos elementos não diagonais da matriz dada. Agora desde que $\left(1 - \frac{2}{n^2 - n}\right) < 1$, segue que $S_k \to 0$ quando $k \to \infty$, e isto signufica que $A \to D$, quando $k \to \infty$. Com isso, acabamos de mostrar que o método de Jacobi é convergente para qualquer matriz real simétrica.

Observe ainda que, na prática, não obtemos, em geral uma matriz diagonal, mas sim uma matriz quase diagonal, ou seja, desde que: $S_{k-1} \leq (n^2 - n)(a_{pq}^{k-1})^2 \leq n^2(a_{pq}^{k-1})^2$, paramos o processo quando $n|a_{pq}^k| < \epsilon$, onde ϵ é uma precisão pré-fixada. A seguir daremos alguns exemplos.

Exemplo 7.9 - *Determinar os auto-valores e correspondentes auto-vetores de:*

$$A = \begin{pmatrix} 7 & 2 \\ 2 & 7 \end{pmatrix} \ ,$$

pelo método de Jacobi.

Solução: Como a matriz é 2×2 para diagonalizar A devemos zerar o elemento $(1,2)$. Assim: $(p,q) = (1,2)$. Temos então que:

$$\phi = \frac{a_{22} - a_{11}}{2a_{12}} = 0 \Rightarrow t = 1 \ .$$

Portanto:

$$\begin{aligned}
c &= \frac{1}{\sqrt{1 + 1^2}} = \frac{\sqrt{2}}{2} = 0.7071 \ , \\
s &= 1 \cdot \frac{1}{\sqrt{2}} = t \times c = \frac{\sqrt{2}}{2} = 0.7071 \ , \\
a''_{11} &= a_{11}c^2 - 2a_{12}sc + a_{22}s^2 \\
&= 7(0.5) - 2(2)(0.7071)(0.7071) + 7(0.5) = 5 \ , \\
a''_{22} &= a_{11}s^2 + 2a_{12}sc + a_{22}c^2 \\
&= 7(0.5) + 2(2)(0.7071)(0.7071) + 7(0.5) = 9 \ ,
\end{aligned}$$

onde utilizamos as fórmulas: (7.7) e (7.8). Assim: $A_1 = \begin{pmatrix} 5 & 0 \\ 0 & 9 \end{pmatrix}$.

Logo os auto-valores de A são: $\lambda_1 = 5; \lambda_2 = 9$ e desde que:

$$V = U_1 = \begin{pmatrix} cos\varphi & sen\varphi \\ -sen\varphi & cos\varphi \end{pmatrix} = \begin{pmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{pmatrix} \ ,$$

os auto-vetores, correspondentes, são:

$$v_1 = \begin{pmatrix} 0.7071 \\ -0.7071 \end{pmatrix} \quad , v_2 = \begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix} \ .$$

Exemplo 7.10 - *Determinar, usando o método de Jacobi, os auto-valores da matriz:*

$$A = \begin{pmatrix} 4 & 2 & 0 \\ 2 & 5 & 3 \\ 0 & 3 & 6 \end{pmatrix} \ .$$

Solução: O maior elemento, em módulo, fora da diagonal principal da matriz $A_1 = A$, é o elemento $a_{23} = a_{32} = 3$. Assim:

$$\phi = \frac{a_{33} - a_{22}}{2a_{23}} = \frac{6-5}{6} = 0.1667 \,.$$

Portanto, $t = 0.8471$, $cos\varphi = c = 0.7630$, $sen\varphi = s = 0.6464$. Como já dissemos podemos ou aplicar as fórmulas: (7.13), (7.14), (7.6) com $j \neq 2,3$ e (7.5) com $i \neq 2,3$, ou simplesmente efetuar o produto $U_1^t A_1 U_1$, para obter A_2, onde:

$$U_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.7630 & 0.6464 \\ 0 & -0.6464 & 0.7630 \end{pmatrix} \Rightarrow A_2 = \begin{pmatrix} 4 & 1.5260 & 1.2928 \\ 1.5260 & 2.4586 & 0 \\ 1.2928 & 0 & 8.5414 \end{pmatrix} .$$

O elemento de maior valor absoluto, na matriz A_2 é $a_{12} = a_{21} = 1.5260$. Assim:

$$\phi = -0.5050, \quad t = -0.6153, \quad c = 0.8517, \quad s = -0.5240 \,.$$

Obtemos, então:

$$U_2 = \begin{pmatrix} 0.8517 & -0.5240 & 0 \\ 0.5240 & 0.8517 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow A_3 = \begin{pmatrix} 4.9387 & 0 & 1.1011 \\ 0 & 1.5197 & -0.6774 \\ 1.1011 & -0.6774 & 8.5414 \end{pmatrix} .$$

Agora $(p,q) = (1,3)$, $\phi = 1.6360$, $t = 0.2814$, $c = 0.9626$, $s = 0.2709$, e com isso obtemos:

$$U_3 = \begin{pmatrix} 0.9626 & 0 & 0.2709 \\ 0 & 1 & 0 \\ -0.2709 & 0 & 0.9626 \end{pmatrix} \Rightarrow A_4 = \begin{pmatrix} 4.6611 & 0.1239 & 0 \\ 0.1239 & 1.5197 & -0.6520 \\ 0 & -0.6520 & 8.8536 \end{pmatrix} .$$

Temos $(p,q) = (2,3)$ e assim afetuando os cálculos segue que: $\phi = -5.6266$, $t = -0.0882$, $c = 0.9961$, $s = -0.0879$. Portanto:

$$U_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.9961 & -0.0879 \\ 0 & -0.0879 & 0.9961 \end{pmatrix} \Rightarrow A_5 = \begin{pmatrix} 4.6228 & 0.1827 & -0.0161 \\ 0.1827 & 1.4621 & 0 \\ -0.0161 & 0 & 8.9081 \end{pmatrix} .$$

Observe que os elementos não diagonais da sequência $A_k \rightarrow 0$, à medida que k aumenta. Assim os elementos diagonais da sequência A_k convergem para os auto-valores de A que são: 1.45163, 4.63951, 8.90885. Uma precisão maior pode ser obtida continuando o processo. Além disso, se desejarmos uma aproximação para os auto-vetores, basta efetuar o produto $U_1 U_2 U_3 U_4$.

7.5.2 Método Cíclico de Jacobi

A procura do elemento de maior módulo, fora da diagonal principal, a cada passo do método de Jacobi, é um processo caro que deve ser evitado. Uma alternativa é percorrer ciclicamente os elementos fora da diagonal principal, por linha, por exemplo. Assim, sucessivamente, zeramos os elementos das posições:

$$\begin{array}{cccc} (1,2) & (1,3) & \ldots & (1,n) \\ & (2,3) & \ldots & (2,n) \\ & & \ldots & \\ & & & (n-1,n) \end{array}$$

escolhendo em cada passo φ tal que $a''_{pq} = 0$. As fórmulas usadas são as mesmas do método de Jacobi. A seguir voltamos à primeira linha, segunda linha, etc, isto é, repetimos o ciclo tantas vezes quantas

forem necessárias até obtermos uma matriz diagonal. Além disso, desde que os elementos não diagonais, a cada passo, decrescem podemos usar uma estratégia conhecida como **Método Cíclico de Jacobi com** *Dados de Entrada*. Tal método consiste em omitir transformações sobre elementos cujo valor, em módulo, é menor que os valores fornecidos como *dados de entrada*. A vantagem deste método é que zeros são criados apenas nas posições onde o valor é em módulo maior que os valores fornecidos nos *dados de entrada*, sem a necessidade de ir zerando todos os elementos. O próximo exemplo ilustra esse método.

Exemplo 7.11 - *Determinar os auto-valores e correspondentes auto-vetores da matriz:*

$$A = \begin{pmatrix} 3 & 0.4 & 5 \\ 0.4 & 4 & 0.1 \\ 5 & 0.1 & -2 \end{pmatrix},$$

usando o método de Jacobi, tomando como dados de entrada para o primeiro e segundo ciclos: 0.5 *e* 0.05*, respectivamente.*

Solução: Para o primeiro ciclo a transformação sobre o elemento $(1,2)$ será omitida pois $|0.4| < 0.5$. Portanto, desde que $|5| > 0.5$, um zero será criado na posição $(1,3)$. Assim, fazendo os cálculos, obtemos:

$$U_1 = \begin{pmatrix} 0.8507 & 0 & -0.5257 \\ 0 & 1 & 0 \\ 0.5257 & 0 & 0.8507 \end{pmatrix} \Rightarrow A_2 = \begin{pmatrix} 6.0902 & 0.3928 & 0 \\ 0.3928 & 4 & -0.1252 \\ 0 & -0.1252 & -5.0902 \end{pmatrix}.$$

A transformação $(2,3)$ será omitida porque $|-0.1252| < 0.5$. Isto completa o primeiro ciclo. Para o segundo ciclo um zero será criado na posição $(1,2)$ porque $|0.3928| > 0.05$. Portanto:

$$U_2 = \begin{pmatrix} 0.9839 & -0.1788 & 0 \\ 0.1788 & 0.9839 & 0 \\ 0 & 0 & 1 \end{pmatrix} \Rightarrow A_3 = \begin{pmatrix} 6.1616 & 0 & -0.0224 \\ 0 & 3.9286 & -0.1232 \\ -0.0224 & -0.1232 & -5.0902 \end{pmatrix}.$$

A transformação $(1,3)$ será omitida pois $|-0.0224| < 0.05$. Finalmente um zero será criado na posição $(2,3)$. Assim:

$$U_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.9999 & 0.0137 \\ 0 & -0.0137 & 0.9999 \end{pmatrix} \Rightarrow A_4 = \begin{pmatrix} 6.1616 & 0.0003 & -0.0224 \\ 0.0003 & 3.9303 & 0 \\ -0.0024 & 0 & -5.0919 \end{pmatrix}.$$

e portanto podemos dizer que os auto-valores de A são aproximadamente iguais a $6.1616, 3.9303$ e -5.0919. Agora, para obtermos os auto-vetores calculamos o produto $U_1U_2U_3$. Fazendo isso, segue que:

$$U_1U_2U_3 = \begin{pmatrix} 0.8370 & -0.1449 & -0.5277 \\ 0.1788 & 0.9838 & 0.0135 \\ 0.5172 & -0.1056 & 0.8439 \end{pmatrix}.$$

Portanto os auto-vetores aproximados de A, correspondentes aos auto-valores aproximados: $6.1616, 3.9303$ e -5.0919, são:

$$\begin{pmatrix} 0.8370 \\ 0.1788 \\ 0.5172 \end{pmatrix}, \quad \begin{pmatrix} -0.1449 \\ 0.9838 \\ -0.1056 \end{pmatrix}, \quad \begin{pmatrix} -0.5277 \\ 0.0135 \\ 0.8439 \end{pmatrix}.$$

Os auto-valores de A são: $6.16161, 3.93029$ e -5.09190.

Observe que os teoremas de Gerschgorin (Teorema 1.10) fornecem ainda um limitante para os erros cometidos nos auto-valores calculados pelo método de Jacobi. No exemplo 7.11, os círculos de Gerschgorin da matriz transformada A_4 são dados por:

$$\begin{array}{rclrcl} a_1 & = & 6.1616 \ , & r_1 & = & 0.0227 \ , \\ a_2 & = & 3.9303 \ , & r_2 & = & 0.0003 \ , \\ a_3 & = & -5.0919 \ , & r_3 & = & 0.0224 \ . \end{array}$$

Estes círculos são isolados e assim existe exatamente um auto-valor em cada círculo. Os auto-valores podem portanto serem estimados por:

$$6.1616 \pm 0.0227 \ , \quad 3.9303 \pm 0.0003 \ , \quad -5.0919 \pm 0.0224$$

De um modo geral, se os elementos não diagonais de uma matriz $n \times n$ simétrica têm módulo não excedendo ϵ então, desde que os círculos de Gerschgorin são isolados, os auto-valores diferem dos elementos da diagonal principal por no máximo $(n-1)\epsilon$.

Exercícios

7.9 - *Determine os auto-valores e auto-vetores das seguintes matrizes:*

$$A = \begin{pmatrix} 10 & -6 & -4 \\ -6 & 11 & 2 \\ -4 & 2 & 6 \end{pmatrix} \ , \quad B = \begin{pmatrix} 2 & 4 & -2 \\ 4 & 2 & 2 \\ -2 & 2 & 5 \end{pmatrix} \ ,$$

usando:

a) *o método de Jacobi,*

b) *o método cíclico de Jacobi,*

c) *o método cíclico de Jacobi, com dados de entrada igual a 10^{-i} para o i-ésimo ciclo.*

7.10 - *Se:*

$$U = \begin{pmatrix} cos\varphi & 0 & sen\varphi \\ 0 & 1 & 0 \\ -sen\varphi & 0 & cos\varphi \end{pmatrix} \ ; \quad A = \begin{pmatrix} 5 & 0 & 1 \\ 0 & -3 & 0.1 \\ 1 & 0.1 & 2 \end{pmatrix} \ ,$$

calcule $U^t AU$, e deduza que se $\phi = -\frac{3}{2}$ então os elementos $(1,3)$ e $(3,1)$ deste produto são iguais a zero. Escreva aproximações para os auto-valores e auto-vetores de A. Use o teorema de Gerschgorin para obter um limite superior do erro nos auto-valores estimados.

7.6 Método de Rutishauser (ou Método LR)

O **método de Rutishauser** ou **Método LR** permite, sob certas condições, determinar todos os auto-valores de uma matriz, sem determinar o polinômio característico.

Seja A uma matriz quadrada de ordem n. O método consiste em construir uma sequência de matrizes $A_1, A_2, \ldots$ do seguinte modo: decompomos $A = A_1$ no produto $L_1 R_1$ onde L_1 é triangular inferior com 1 na diagonal e R_1 é triangular superior. (Decomposição LU, Capítulo 4). Então, $A_1 = L_1 R_1$. Agora,

multiplicamos as duas matrizes na ordem inversa e formamos a matriz $A_2 = R_1L_1$, e decompomos, a seguir, a matriz A_2 no produto de duas matrizes triangulares L_2 e R_2 e assim por diante. Então temos:

$$\begin{array}{l} A_1 = A = L_1R_1 \\ A_2 = R_1L_1 = L_2R_2 \\ A_3 = R_2L_2 = L_3R_3 \\ \vdots \\ A_k = R_{k-1}L_{k-1} = L_kR_k \\ \vdots \end{array}$$

Observações:

1) Pode-se provar que: Se os auto-valores de A são distintos a sequência $\{A_k\}$ converge para uma matriz triangular superior R.

2) As matrizes A e R são matrizes similares. De fato, temos: $A_1 = L_1R_1 \Rightarrow L_1^{-1}A_1 = R_1$, então:

$$A_2 = R_1L_1 = L_1^{-1}AL_1 \ ,$$

desde que $A_1 = A$. Portanto A_2 é similar a A. De $A_2 = L_2R_2 \Rightarrow L_2^{-1}A_2 = R_2$, então:

$$A_3 = R_2L_2 = L_2^{-1}A_2L_2 = L_2^{-1}L_1^{-1}AL_1L_2 \ ,$$

e portanto A_3 é similar a A. De um modo geral, obtemos:

$$A_k = R_{k-1}L_{k-1} = \underbrace{L_{k-1}^{-1} \dots L_1^{-1}}_{L^{-1}} A \underbrace{L_1 \dots L_{k-1}}_{L} \ .$$

Portanto A_k é similar a A. Logo possuem o mesmo polinômio característico. Portanto possuem os mesmos auto-valores.

3) Os elementos diagonais da matriz A_k são os auto-valores procurados.

4) O processo termina quando o elemento de maior valor absoluto da matriz A_k, (abaixo da diagonal principal), for menor que ϵ, onde ϵ é uma precisão pré-fixada.

Exemplo 7.12 - *Calcular os auto-valores de:*

$$A = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

pelo método de Rutishauser com precisão de 10^{-2}.

Solução: Temos:

$$\begin{aligned}
A_1 &= A = \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0.5 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 1 \\ & 1 & 0 \\ & & 0.5 \end{pmatrix} = L_1 U_1 \ , \\
A_2 &= U_1 L_1 \ = \begin{pmatrix} 2.5 & 0 & 1 \\ 0 & 1 & 0 \\ 0.25 & 0 & 0.5 \end{pmatrix} \\
&= \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0.1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2.5 & 0 & 1 \\ & 1 & 0 \\ & & 0.4 \end{pmatrix} = L_2 U_2 \ , \\
A_3 &= U_2 L_2 \ = \begin{pmatrix} 2.6 & 0 & 1 \\ 0 & 1 & 0 \\ 0.04 & 0 & 0.4 \end{pmatrix} \\
&= \begin{pmatrix} 1 & & \\ 0 & 1 & \\ 0.0154 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2.6 & 0 & 1 \\ & 1 & 0 \\ & & 0.3846 \end{pmatrix} = L_3 U_3 \ , \\
A_4 &= U_3 L_3 \ = \begin{pmatrix} 2.6154 & 0 & 1 \\ 0 & 1 & 0 \\ 0.00592 & 0 & 0.3846 \end{pmatrix} .
\end{aligned}$$

Como os elementos abaixo da diagonal principal de A_4 são, em módulo menor que $10^{-2} \Rightarrow A_4 \simeq R$. Assim, os auto-valores de A são:

$$\begin{aligned}
&\lambda_1 \simeq 2.6154 \ , \\
&\lambda_2 = 1 \ , \\
&\lambda_3 \simeq 0.3846 \ , \quad \text{com} \quad \epsilon < 10^{-2}.
\end{aligned}$$

Observe que os auto-valores de A são: 2.618034, 1 e 0.381966.

O método de Rutishauser permite obter também os auto-vetores. Entretanto o cálculo dos auto-vetores, por este método, é um tanto trabalhoso e assim será omitido. O leitor interessado pode encontrar a descrição do método, por exemplo em [Fox, 19..].

Exercícios

7.11 - *Usando o método LR, determine os auto-valores das matrizes:*

$$A \ = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 2 & 2 \\ 1 & 2 & 5 \end{pmatrix} \ , \ B \ = \begin{pmatrix} 5 & 1 & 0 \\ -1 & 3 & 1 \\ -2 & 1 & 10 \end{pmatrix} \ ,$$

com precisão de 10^{-2}.

7.12 - *Considere a matriz:*

$$A = \begin{pmatrix} 5 & 0 & 1 \\ 0 & 1 & 0 \\ 5 & 0 & 1 \end{pmatrix} .$$

Usando o método LR, uma única vez, isto é, até determinar A_2*, é possível estimar os auto-valores de* A*?*

7.7 Método de Francis (ou Método QR)

O **método de Francis** ou **Método QR** determina todos os auto-valores de uma matriz, sem determinar o polinômio característico.

Seja A uma matriz quadrada de ordem n. O método consiste em construir uma sequência de matrizes $A_1, A_2, \ldots$ do seguinte modo: decompomos $A = A_1$ no produto $Q_1 R_1$ onde Q_1 é ortogonal e R_1 é triangular superior. Então, $A_1 = Q_1 R_1$. Agora, multiplicamos as duas matrizes na ordem inversa e formamos a matriz $A_2 = R_1 Q_1$, e decompomos, a seguir, a matriz A_2 no produto $Q_2 R_2$ e assim por diante. Então temos:

$$\begin{array}{lcl} A_1 & = & A \; = \; Q_1 R_1 \; , \\ A_2 & = & R_1 Q_1 \; = \; Q_2 R_2 \; , \\ \vdots & & \\ A_k & = & R_{k-1} Q_{k-1} \; = \; Q_k R_k \; \vdots \end{array}$$

Observações:

a) Essa decomposição tem a vantagem, em relação ao método LR, de sempre existir. Além disso, se A_s é real então Q_s e R_s são reais.

b) A sequência A_k converge para uma matriz triangular superior em cuja diagonal encontram-se os auto-valores da matriz A.

c) A matriz A_k é similar a matriz A. De fato, temos: $A_1 = Q_1 R_1 \;\Rightarrow\; Q_1^{-1} A_1 = R_1$, então:

$$A_2 \; = \; R_1 Q_1 \; = \; Q_1^{-1} A Q_1$$

Portanto, desde que $A_1 = A$, temos que: A_2 e A são similares. De um modo geral, obtemos:

$$A_{k+1} \; = \; R_k Q_k = \underbrace{Q_k^{-1} Q_{k-1}^{-1} \ldots Q_1^{-1}}_{Q^{-1}} A_1 \underbrace{Q_1 \ldots Q_{k-1} Q_k}_{Q}$$

Portanto A_{k+1} é similar a A. Logo possuem o mesmo polinômio característico. Portanto possuem os mesmos auto-valores.

d) Os elementos diagonais da matriz A_k são os auto-valores procurados.

e) O processo termina quando o elemento de maior valor absoluto da matriz A_k, (abaixo da diagonal principal), for menor que ϵ, onde ϵ é uma precisão pré-fixada.

Em cada passo do método QR, devemos determinar matrizes Q_k e R_k onde Q_k é matriz ortogonal e R_k é matriz triangular superior. Essa decomposição pode ser obtida utilizando transformações ortogonais da forma (1.23). A seguir mostramos como isso pode ser feito.

Seja A uma matriz que desejamos decompor no produto QR. Para zerar o elemento a_{21}, fazemos o produto $U_1 A$ e com isso obtemos uma matriz $A^{(1)}$; para zerar o elemento a_{31} fazemos o produto $U_2 A^{(1)}$ e assim obtemos uma matriz $A^{(2)}$, e assim sucessivamente, isto é, procedemos coluna por coluna até zerarmos todos os elementos abaixo da diagonal principal. O produto das matrizes $U_1^t U_2^t \ldots$ fornece a matriz Q_1.

Considere então o produto $U_1 A$, onde U_1 é dada por (1.23). O elemento a'_{qp} é dado por:

$$a'_{qp} \;=\; -sen\;\varphi a_{pp} + cos\;\varphi a_{qp}\;, \tag{7.16}$$

e queremos $a'_{qp} = 0$. Assim, o que desejamos é

$$\begin{aligned}
- \quad a_{pp}\sqrt{1-cos^2\varphi} + cos\varphi a_{qp} &= 0\\
\Rightarrow \quad a_{pp}\sqrt{1-cos^2\varphi} &= a_{qp}cos\varphi\\
\Rightarrow \quad a_{pp}^2(1-cos^2\varphi) &= a_{qp}^2 cos^2\varphi\\
\Rightarrow \quad (a_{pp}^2+a_{qp}^2)cos^2\varphi &= a_{pp}^2\\
\Rightarrow \quad cos\varphi &= \frac{a_{pp}}{\sqrt{a_{pp}^2+a_{qp}^2}}\,.
\end{aligned}$$

Por outro lado, igualando (7.18) a zero, segue que:

$$sen\varphi \;=\; \frac{a_{qp}\;cos\;\varphi}{a_{pp}} \;\Rightarrow\; sen\varphi \;=\; \frac{a_{qp}}{\sqrt{a_{pp}^2+a_{qp}^2}}\,.$$

Para melhor entendimento do método, considere uma matriz de ordem 3. Para reduzí-la a forma triangular devemos zerar os elementos a_{21}, a_{31} e a_{32}. Assim, fazendo $c = cos\varphi$ e $s = sen\varphi$, segue que:

1) para zerar o elemento a_{21}, efetuamos o produto:

$$\underbrace{\begin{pmatrix} c & s & 0\\ -s & c & 0\\ 0 & 0 & 1\end{pmatrix}}_{U_1}\begin{pmatrix} a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix} = \begin{pmatrix} a'_{11} & a'_{12} & a'_{13}\\ 0 & a'_{22} & a'_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix}\,,$$

e desde que queremos $a_{21} \;=\; 0$ devemos ter:

$$-s\;a_{11} + c\;a_{21} = 0,\;\text{onde}$$

$$s \;=\; \frac{a_{21}}{\sqrt{a_{11}^2+a_{21}^2}} \quad e \quad c \;=\; \frac{a_{11}}{\sqrt{a_{11}^2+a_{21}^2}}\,.$$

2) para zerar o elemento a_{31}, efetuamos o produto:

$$\underbrace{\begin{pmatrix} c & 0 & s\\ 0 & 1 & 0\\ -s & 0 & c\end{pmatrix}}_{U_2}\begin{pmatrix} a'_{11} & a'_{12} & a'_{13}\\ 0 & a'_{22} & a'_{23}\\ a_{31} & a_{32} & a_{33}\end{pmatrix} = \begin{pmatrix} a''_{11} & a''_{12} & a''_{13}\\ 0 & a'_{22} & a'_{23}\\ 0 & a''_{32} & a''_{33}\end{pmatrix}$$

e desde que queremos $a_{31} = 0$, devemos ter:

$$-\;s\;a'_{11} + c\;a_{31} = 0,\;\text{onde}$$

$$s \;=\; \frac{a_{31}}{\sqrt{a'^2_{11}+a_{31}^2}} \quad \text{e} \quad c \;=\; \frac{a'_{11}}{\sqrt{a'^2_{11}+a_{31}^2}}\,.$$

3) para zerar o elemento a_{32}, efetuamos o produto:

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & c & s \\ 0 & -s & c \end{pmatrix}}_{U_3} \begin{pmatrix} a''_{11} & a''_{12} & a''_{13} \\ 0 & a'_{22} & a'_{23} \\ 0 & a''_{32} & a''_{33} \end{pmatrix} = \begin{pmatrix} a''_{11} & a''_{12} & a''_{13} \\ 0 & a''_{22} & a''_{23} \\ 0 & 0 & a''_{33} \end{pmatrix}$$

e desde que queremos $a''_{32} = 0$, devemos ter:

$$- s\ a'_{22} + c\ a''_{32} \ = \ 0$$

$$\Rightarrow \ s \ = \ \frac{a''_{32}}{\sqrt{a'^2_{22} + a''^2_{32}}} \qquad \text{e} \qquad c \ = \ \frac{a'_{22}}{\sqrt{a'^2_{22} + a''^2_{32}}} \ .$$

Assim, obtemos:

$$U_3 U_2 U_1 A = R_1 \ \Rightarrow \ A = \underbrace{U_1^t U_2^t U_3^t}_{Q_1} R_1 \ .$$

O produto $R_1 Q_1 \ = \ R_1 U_1^t U_2^t U_3^t$ é obtido por sucessivas pré-multiplicações de R com as matrizes $U_k^t, \ k = 1, 2, \ldots$.

Exemplo 7.13 - *Determinar os auto-valores da matriz:*

$$A \ = \ \begin{pmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

pelo método de Francis; com $\epsilon < 10^{-2}$.

Solução: Como $a_{21} = 0$, devemos zerar apenas o elemento a_{31}. Assim $U_1 = I$ e para obtermos U_2, fazemos:

$$s \ = \ \frac{a_{31}}{\sqrt{a_{11}^2 + a_{31}^2}} \ \Rightarrow \ s = \frac{1}{\sqrt{2^2 + 1^2}} \ = \ \frac{1}{\sqrt{5}} \ = \ 0.4472 \ ,$$

$$c \ = \ \frac{a_{11}}{\sqrt{a_{11}^2 + a_{31}^2}} \ \Rightarrow \ c = \frac{2}{\sqrt{5}} \ = \ 0.8944 \ .$$

Assim :

$$U_2 \ = \ \begin{pmatrix} 0.8944 & 0 & 0.4472 \\ 0 & 1 & 0 \\ -0.4472 & 0 & 0.8944 \end{pmatrix} .$$

Portanto:

$$U_2 U_1 A \ = \ U_2 I A \ = \ U_2 A \ = \underbrace{\begin{pmatrix} 2.2360 & 0 & 1.3416 \\ 0 & 1 & 0 \\ 0 & 0 & 0.4472 \end{pmatrix}}_{R_1} .$$

Desde que $a_{32} = 0 \ \Rightarrow \ U_3 = I$. Assim: $U_3 U_2 U_1 \ = \ U_2$ e $U_2^{-1} \ = \ U_2^t$. Portanto:

$$A_1 = A = \underbrace{\begin{pmatrix} 0.8944 & 0 & -0.4472 \\ 0 & 1 & 0 \\ 0.4472 & 0 & 0.8944 \end{pmatrix}}_{U_2^t} \begin{pmatrix} 2.2360 & 0 & 1.3416 \\ 0 & 1 & 0 \\ 0 & 0 & 0.4472 \end{pmatrix} \ = \ Q_1 R_1 \ .$$

Agora:

$$A_2 = R_1 Q_1 = \begin{pmatrix} 2.5998 & 0 & 0.2000 \\ 0 & 1 & 0 \\ 0.2000 & 0 & 0.4000 \end{pmatrix} .$$

Aplicando novamente o processo, temos que: $U_1 = U_3 = I$. Devemos então determinar U_2. Assim:

$$\begin{aligned} s &= \frac{0.2000}{\sqrt{(2.5998)^2 + (0.2000)^2}} = 0.0767 , \\ c &= \frac{2.5998}{\sqrt{(2.5998)^2 + (0.2000)^2}} = 0.9971 . \end{aligned}$$

Portanto:

$$U_2 = \begin{pmatrix} 0.9971 & 0 & 0.0767 \\ 0 & 1 & 0 \\ -0.0767 & 0 & 0.9971 \end{pmatrix} ,$$

e assim:

$$U_2 A_2 = \begin{pmatrix} 2.6076 & 0 & 0.2301 \\ 0 & 1 & 0 \\ 0 & 0 & 0.3935 \end{pmatrix} = R_2$$

Logo:

$$A_2 = \underbrace{\begin{pmatrix} 0.9971 & 0 & -0.0767 \\ 0 & 1 & 0 \\ 0.0767 & 0 & 0.9971 \end{pmatrix}}_{U_2^t} \begin{pmatrix} 2.6076 & 0 & 0.2301 \\ 0 & 1 & 0 \\ 0 & 0 & 0.3835 \end{pmatrix} = Q_2 \; R_2$$

Finalmente,

$$A_3 = R_2 \; Q_2 = \begin{pmatrix} 2.6177 & 0 & 0.0294 \\ 0 & 1 & 0 \\ 0.0094 & 0 & 0.3824 \end{pmatrix} .$$

Desde que o maior elemento, em valor absoluto, abaixo da diagonal principal é menor do que 10^{-2}, temos que os valores aproximados dos auto-valores de A são: 2.6177, 1 e 0.3824. Observe que os auto-valores de A são: 2.618034, 1 e 0.381966.

O método QR permite obter também os auto-vetores. Como no método LR o cálculo dos auto-vetores é trabalhoso por este método e assim será omitido. O leitor interessado pode encontrar a descrição do método, por exemplo em [Fox, 19..].

Exercícios

7.13 - *Usando o método QR, determinar todos os auto-valores das matrizes:*

$$A = \begin{pmatrix} 4 & 4 & -3 \\ 0 & 8 & 1 \\ 0 & 2 & -1 \end{pmatrix} , \quad B = \begin{pmatrix} 12 & 3 & 1 \\ -9 & -2 & -3 \\ 14 & 6 & 2 \end{pmatrix} ,$$

com precisão de 10^{-2}.

7.14 - *Usando o método QR, uma única vez, na matriz:*

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 2 & 0 & 1 \\ 2 & 1 & -1 \end{pmatrix} ,$$

é possível estimar seus auto-valores? (Use aritmética exata).

7.8 Exercícios Complementares

7.15 - *Para cada uma das matrizes:*

$$A = \begin{pmatrix} -2 & 5 \\ 1 & -3 \end{pmatrix} , \quad A = \begin{pmatrix} 1 & 4 & 3 \\ 0 & 3 & 1 \\ 0 & 2 & -1 \end{pmatrix} ,$$

encontre um polinômio que tenha a matriz como raiz.

7.16 - *Sabendo que uma matriz de ordem 3 tem como auto-valores* $\lambda_1 = -1$, $\lambda_2 = 2$, $\lambda_3 = 3$.

a) *Qual é o polinômio característico de A?*

b) *Quanto vale* $tr(A^2)$*?*

c) *Quais são os auto-valores de* A^{-1}*?*

d) *A matriz A é uma matriz singular? Por quê?*

7.17 - *Seja A uma matriz quadrada de ordem n e sejam* $\lambda_1, \lambda_2, \cdots, \lambda_n$ *seus auto-valores. Quais são os auto-valores de* $A - qI$ *onde q é uma constante e I é a matriz identidade?*

7.18 - *Mostre que se v é auto-vetor de A e de B então v é auto-vetor de* $\alpha A + \beta B$*, onde* α, β *são escalares quaisquer.*

7.19 - *Mostre que uma matriz A e sua transposta* A^t *possuem o mesmo polinômio característico.*

7.20 - *Considere a matriz:*

$$A = \begin{pmatrix} 1 & 3 & -1 \\ 0 & 0 & 2 \\ -1 & 1 & 0 \end{pmatrix} .$$

Verifique, através do método de Leverrier, que seu polinômio característico é dado por:

$$P(\lambda) = -\lambda^3 + \lambda^2 + 3\lambda - 8 .$$

7.21 - *Seja a matriz:*

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 3 & 1 \\ 2 & 1 & 2 \end{pmatrix} .$$

a) *Verifique pelo método de Leverrier-Faddeev que seu polinômio característico é dado por:*

$$P(\lambda) = (-1)^3(\lambda^3 - 6\lambda^2 + 6\lambda + 7) .$$

b) *Determine por método numérico a sua escolha o único auto-valor real negativo de* A *com precisão de* 10^{-2}.

c) *Usando os resultados obtidos em a) e b) calcule o auto-vetor correspondente .*

d) *Usando* **a)** *obtenha a inversa de* A.

7.22 - *Usando o método das potências determine, com precisão de* 10^{-3}*, o auto-valor de maior valor absoluto, e seu correspondente auto-vetor, para cada uma das seguintes matrizes:*

$$A = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 3 & 2 \\ 1 & 2 & 3 \end{pmatrix} \quad , \quad B = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \; .$$

7.23 - *Considere a matriz:*

$$A = \begin{pmatrix} 1 & -1 & 2 \\ -1 & 1 & 2 \\ 2 & -2 & 8 \end{pmatrix} \; .$$

a) *Pelo método das potências calcule o auto-valor de maior valor absoluto de* A *e seu correspondente auto-vetor.*

b) *Obtenha o polinômio caracteristico de* A *pelo método de Leverrier-Faddeev.*

c *Determine os demais auto-valores de* A.

d) *Obtenha o auto-vetor correspondente ao auto-valor* λ_2 *pelo processo de Leverrier - Faddeev. Suponha* $|\lambda_1| > |\lambda_2| > |\lambda_3|$.

7.24 - *Determinar o auto-valor de maior valor absoluto da matriz:*

$$A = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 5 & 1 \\ 2 & 1 & 6 \end{pmatrix} \; .$$

usando o método das potências. Use como vetor inicial $y_0 = (8/9, \; 8/9, \; 1)^t$*. Dê seu valor aproximado após três iterações.*

7.25 - *Considere as matrizes:*

$$A = \begin{pmatrix} 1 & 3 \\ -1 & 5 \end{pmatrix} ; \quad B = \begin{pmatrix} 1 & -3 & 3 \\ 3 & -5 & 3 \\ 6 & -6 & 4 \end{pmatrix} \; .$$

Para cada uma delas:

a) *calcule* $P(\lambda)$ *e suas raízes algebricamente.*

b) *calcule* $P(\lambda)$ *pelo método de Leverrier.*

c) *calcule os auto-valores e auto-vetores pelo método de Leverrier-Faddeev.*

d) *calcule os auto-valores pelo método de potências.*

e) *calcule os auto-valores pelo método LR.*

f) *calcule os auto-valores pelo método QR.*

7.26 - *Matrizes do tipo:*

$$\begin{pmatrix} x_0 & x_1 & x_2 \\ x_2 & x_0 & x_1 \\ x_1 & x_2 & x_0 \end{pmatrix} ,$$

são chamadas **matrizes circulantes**. *Determine todos os auto-valores e correspondentes auto-vetores da matriz circulante onde* $x_0 = 9, x_1 = 2$ *e* $x_3 = 1$, *utilizando para isso método numérico a sua escolha.*

7.27 - *Localizar, usando o teorema de Gerschgorin, os auto-valores de:*

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 0 & 3 & 1 \\ 2 & 1 & 2 \end{pmatrix} , \quad B = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & -1 & 0 \end{pmatrix} .$$

7.28 - *Considere a matriz:*

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} .$$

Determine os auto-valores de A, usando:

a) *o método clássico de Jacobi,*

b) *o método cíclico de Jacobi,*

c) *o método cíclico de Jacobi, com dados de entrada igual a* 5×10^{-i} *para o i-ésimo ciclo.*

d) *Use o teorema de Gerschgorin para obter um limite superior do erro nos auto-valores estimados.*

7.29 - *Considere a matrizes:*

$$A = \begin{pmatrix} 10 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} , \quad B = \begin{pmatrix} 100 & 0 & 99 \\ 0 & 20 & 0 \\ 99 & 0 & 101 \end{pmatrix} .$$

a) *Caso haja convergência pelo método de Rutishauser, o que se deve esperar?*

b) *Determine os auto-valores usando o método de Rutishauser. Use com processo de parada* $\epsilon = 10^{-2}$, *ou número máximo de iterações = 3.*

7.30 *Considere a matriz:*

$$A = \begin{pmatrix} 4 & 1 & 1 \\ 2 & 4 & 1 \\ 0 & 1 & 4 \end{pmatrix} .$$

Determine os auto-valores de A, com precisão de 10^{-2}, *usando:*

a) *o método LR,*

b) *o método QR.*

7.9 Problemas Aplicados e Projetos

7.1 - *Considere o movimento horizontal do conjunto massa mola mostrado na Figura 7.3.*

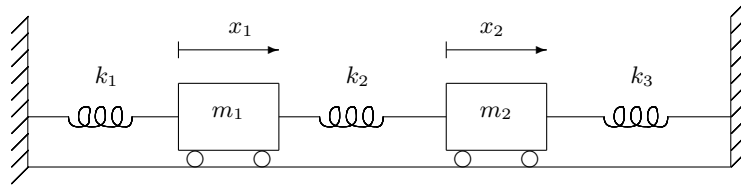



Figura 7.3

As deflexões horizontais x_1 e x_2 são medidas relativamente à posição de equilíbrio estático. As molas possuem rigidez k_1, k_2 e k_3, que são as forças requeridas para estender ou comprimir cada mola de uma unidade de comprimento.
As equações de movimento são:

$$m_1 \frac{d^2 x_1}{dt^2} = -k_1 x_1 + k_2(x_2 - x_1)$$

$$m_2 \frac{d^2 x_2}{dt^2} = k_2(x_1 - x_2) + k_3 x_2$$

a) *Se $x = (x_1, x_2)^t$ é o vetor deflexão então podemos reescrever as equações acima na forma:*

$$\frac{d^2 x}{dt^2} = Ax$$

b) *Mostre que a substituição:*

$$x = ve^{iwt}$$

onde v é um vetor do R^2 , $e^{iwt} = coswt + isenwt$ com $i = \sqrt{-1}$, leva ao problema de auto-valores : $Av = \lambda v$ onde $\lambda = -w^2$. Os possíveis valores que w pode assumir são as frequências naturais de vibração do sistema.

c) *Se $k_1 = k_2 = k_3 = 1kg/s^2$ e $m_1 = m_2 = 1kg$ determine os auto-valores e auto-vetores de A, por método numérico 'a sua escolha.*

7.2 - *Considere o seguinte sistema de equações diferenciais, com coeficientes constantes:*

$$\begin{cases} \frac{dy_1}{dx} = f_1(x, y_1, y_2, \ldots, y_n) \\ \frac{dy_2}{dx} = f_2(x, y_1, y_2, \ldots, y_n) \\ \frac{dy_3}{dx} = f_3(x, y_1, y_2, \ldots, y_n) \\ \vdots \\ \frac{dy_n}{dx} = f_3(x, y_1, y_2, \ldots, y_n) \end{cases} , \qquad (7.17)$$

Se escrevermos (7.17), na forma:

$$Y'(t) = AY(t) ,$$

então a solução geral do sistema é dado por:

$$Y(t) \; = \; \sum_{k=1}^{n} c_k e^{\lambda_k t} v_k \; ,$$

onde: c_k *são constantes arbitrárias,* λ_k *são os auto-valores de* A *e* v_k *seus correspondentes auto-vetores.*

Considere os sistemas:

$$(I) \left\{ \begin{array}{lcl} \dfrac{dy_1}{dx} & = & 10y_1 \\ \dfrac{dy_2}{dx} & = & y_1 - 3y_2 - 7y_3 \\ \dfrac{dy_3}{dx} & = & 2y_2 + 6y_3 \end{array} \right. , \quad (II) \left\{ \begin{array}{lcl} \dfrac{dy_1}{dx} & = & -10y_1 - 7y_2 + 7y_3 \\ \dfrac{dy_2}{dx} & = & 5y_1 + 5y_2 - 4y_3 \\ \dfrac{dy_3}{dx} & = & -7y_1 - 5y_2 + 6y_3 \end{array} \right. ,$$

Determine a solução geral destes sistemas, usando um método numérico à sua escolha para determinar todos os auto-valores e auto-vetores. Cuidado!! O sistema (II) possui auto-valores iguais em módulo.

7.3 - *A curvatura de uma coluna delgada sujeita a uma carga* P *pode ser modelada por:*

$$\frac{d^2y}{dx^2} \; = \; \frac{M}{EI} \; , \tag{7.18}$$

onde $\frac{d^2y}{dx^2}$ *especifica a curvatura,* M *é o momento de curvatura,* E *é o módulo de elasticidade, e* I *é o momento de inércia da seção transversal sobre o eixo neutro. Considerando o corpo livre na Figura 7.4-b é claro que o momento de curvatura em* x *é* $M = -Py$*. Substituindo esse valor na equação (7.18) resulta:*

$$\frac{d^2y}{dx^2} + p^2 y \; = \; 0 \; , \tag{7.19}$$

onde

$$p^2 = \frac{P}{EI} \; . \tag{7.20}$$

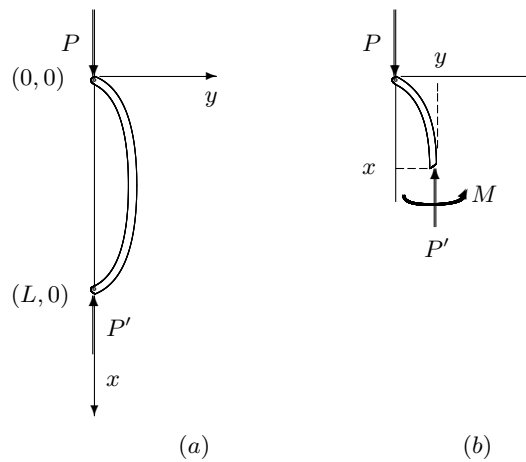



Figura 7.4

Para o sistema na Figura 7.4, sujeita às condições de contorno $y(0) = y(L) = 0$, a solução geral da equação (7.19) é:

$$y = A \; sen \; px + B \; cos \; px \; ,$$

onde A e B são constantes arbitrárias que devem ser obtidas usando-se as condições de contorno. Mostre que de $y(0) = 0$ obtêm-se $B = 0$ e de $y(L) = 0$ obtêm-se $A \; sen \; pL = 0$. Mas desde que $A = 0$ representa a solução trivial, concluímos que $sen \; pL = 0$. Assim, para que esta última igualdade seja válida devemos ter:

$$pL = n\pi \; , \quad n = 1, 2, \ldots \; . \tag{7.21}$$

Portanto, existe um número infinito de valores que satisfazem as condições de contorno. A equação (7.21) pode ser resolvida para:

$$p = \frac{n\pi}{L} \; , \quad n = 1, 2, 3, \ldots \; , \tag{7.22}$$

os quais são os auto-valores para a coluna. Cada auto-valor corresponde ao modo nos quais a coluna curva-se. Combinando as equações (7.20) e (7.22), segue que:

$$P = \frac{n^2\pi^2 EI}{L^2} \; , \quad n = 1, 2, 3, \ldots \; .$$

Isto pode ser entendido como uma deformação da carga porque elas representam os níveis nos quais as colunas movimentam-se em cada deformação sucessiva. Na prática, em geral, o auto-valor correspondente a $n = 1$ é o de interesse porque a quebra usualmente ocorre quando a primeira coluna se deforma. Assim, a carga crítica pode ser definida como:

$$P_{crit.} = \frac{\pi^2 EI}{L^2} \; .$$

Uma carga sobre uma coluna de madeira possui as seguintes características: $E = 10^10 \; Pa$, $I = 1.25 \times 10^{-5} m^4$, e $L = 3m$. Determine o oito primeiros auto-valores, isto é, os auto-valores correspondente a $n = 1, 2 \ldots, 8$ e suas correspondentes deformações das cargas. Qual o valor obtido para a carga crítica?

7.4 - *No problema 7.3 foi razoavelmente fácil obter os auto-valores pois era conhecida a expressão analítica da solução, o que em geral não acontece na prática. Assim, podemos obter os auto-valores de (7.19), substituindo a derivada segunda pela diferença dividida central, isto é, substituindo $\frac{d^2y}{dx^2}$ por:*

$$\frac{y_{i+1} - 2y_i + y_{i-!}}{h^2} \; .$$

Fazendo isso, podemos escrever (7.20) como:

$$\frac{y_{i+1} - 2y_i + y_{i-!}}{h^2} + p^2 y_i = 0 \; ,$$

ou ainda:

$$y_{i+1} - (2 - h^2 \; p^2) y_i + y_{i-!} = 0 \; .$$

Escrevendo esta equação para uma série de nós ao longo do eixo da coluna, obtêm-se um sistema de equações homogêneas. Por exemplo, se a coluna é dividida em cinco segmentos (isto é, quatro nós interiores), o resultado é:

$$\begin{pmatrix} (2 - h^2p^2) & -1 & & \\ -1 & (2 - h^2p^2) & -1 & \\ & -1 & (2 - h^2p^2) & -1 \\ & & -1 & (2 - h^2p^2) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = 0 \; .$$

Considerando os mesmos dados do problema 7.3 determine os auto-valores para os casos: **a)** *um,* **b)** *dois,* **c)** *três e* **d)** *quatro nós interiores, usando método numérico à sua escolha. Lembre-se: desde que* $L = 3$*, segue que para um nó interior* $h = \frac{3}{2}$*, para dois nós interiores* $h = \frac{3}{3}$*, etc.*

7.5 - *No probema 7.4, para três nós interiores você obteve o seguinte sistema:*

$$\begin{pmatrix} (2-0.5625p^2) & -1 & \\ -1 & (2-0.5625p^2) & -1 \\ & -1 & (2-0.5625p^2) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = 0 \ .$$

a) *Mostre que dividindo cada equação por* h^2*, obtêm-se:*

$$(A - \lambda I) = \begin{pmatrix} 3.556-\lambda & -1.778 & \\ -1.778 & 3.556-\lambda & -1.778 \\ & -1.778 & 3.556-\lambda \end{pmatrix}$$

onde $\lambda = p^2$*, e que a expansão do determinante fornece:*

$$P(\lambda) = -\lambda^3 + 10.667\lambda^2 - 31.607\lambda + 22.487 \ .$$

b) *Mostre que o mesmo polinômio pode ser obtido aplicando-se o método de Leverrier-Faddeev à matriz:*

$$B = \begin{pmatrix} 3.556 & -1.778 & \\ -1.778 & 3.556 & -1.778 \\ & -1.778 & 3.556 \end{pmatrix}$$

c) *Usando o polinômio característico obtido em* **b)**, *determine os auto-valores de* B*, usando método numérico à sua escolha.*

d) *Usando o método de Jacobi, determine os auto-valores e auto-vetores da matriz* B*.*

Capítulo 8

Aproximação de Funções: Método dos Mínimos Quadrados

8.1 Introdução

O objetivo do presente capítulo consiste em resolver o seguinte problema: aproximar uma função $y = f(x)$ (real e de variável real) por uma função $F(x)$ que seja combinação linear de funções conhecidas, isto é:

$$f(x) \simeq a_0 \ g_0(x) \ + \ a_1 \ g_1(x) \ + \ \ldots \ + \ a_m \ g_m(x) \ = \ F(x) \ .$$

de tal modo que a distância de $f(x)$ a $F(x)$ seja a menor possível.

A substituição de $f(x)$ por uma função $F(x)$ é indicada quando o uso da função dada oferece alguns inconvenientes, tais como:

a) $f(x)$ é definida através de processos não-finitos como integrais, soma de séries, etc;

b) $f(x)$ é conhecida através de pares de pontos, obtidos em experimentos, e desejamos substituí-la por uma função cujo gráfico se ajuste aos pontos observados;

que podem ser afastados através de uma escolha apropriada da função $F(x)$.

Antes de descrevermos o método do mínimos quadrados relembremos alguns conceitos básicos.

Sabemos da geometria plana euclidiana que: dados uma reta r e um ponto P fora dela, o ponto da reta r mais próximo de P é o único ponto Q tal que PQ é ortogonal a r.

O mesmo acontece na geometria euclidiana sólida, isto é: dados um plano α e um ponto P fora dele, o ponto de α mais próximo de P é o pé da perpendicular traçada de P a α.

Como generalizar tal idéia a um espaço euclidiano E qualquer? O problema que devemos resolver agora é: dados uma função $f(x) \in E$ e um sub-espaço E' de E qual deve ser a função $F(x) \in E'$, tal que:

$$\| f(x) - F(x) \| < \| f(x) - Q(x) \|,$$

para qualquer que seja $Q(x) \in E'$, $Q(x) \neq F(x)$?

Este problema pode ser resolvido através da aproximação de $f(x)$ por $F(x)$ pelo **Método dos Mínimos Quadrados**, como veremos nesse capítulo.

8.2 Aproximação Polinomial

Vamos tratar aqui da aproximação de uma função $y = f(x)$ por um polinômio de um certo grau m, isto é, $F(x) = P_m(x)$, tanto no caso em que $f(x) \in C[a,b]$, onde C[a,b] é o espaço vetorial das funções contínuas reais definidas no intervalo fechado e limitado [a,b] (*caso contínuo*), como no caso onde $f(x)$ é dada por pares de pontos (*caso discreto*).

8.2.1 Caso Contínuo

Consideremos uma função $f(x) \in C[a,b]$.

Inicialmente analisaremos o problema considerando que o polinômio a ser determinado seja escrito em relação à base canônica e a seguir que ele seja escrito em relação a uma base ortonormal. Assim:

Representação na Base Canônica

Desejamos aproximar $f(x),\ x \in [a,b]$, por um polinômio de grau no máximo m, isto é,

$$f(x) \simeq a_0 \ + \ a_1\ x + \ldots + a_m\ x^m \ = \ P_m(x),$$

de tal modo que a distância de f a P_m seja mínima.

Observe que neste caso: $g_0(x) = 1,\ g_1(x) = x,\ \ldots,\ g_m(x) = x^m$ são funções conhecidas.

Assim, o polinômio (a coeficientes reais), $P_m(x)$, deve ser tal que:

$$dist\,(f, P_m) \quad = \ \text{mínima}.$$

Usando a definição de distância, dada anteriormente, (Capítulo 1), temos:

$$\begin{aligned} dist\,(f, P_m) \quad &= \quad \|f - P_m\| \ = \ [(f - P_m, f - P_m)]^{1/2} \\ &= \quad \left[\int_a^b (f(x) \ - \ P_m(x))^2 \ dx\right]^{1/2} \ = \ \|f - P_m\|^2\ . \end{aligned}$$

Assim, o que desejamos é obter:

$$Q \ = \ \|f - P_m\|^2 \quad = \ \text{mínima} \ \ ,$$

(daí a justificativa para o nome **mínimos quadrados**).

Precisamos, então, determinar na classe de todos os polinômios de grau menor ou igual a m aquele que minimize:

$$Q = \| f - P_m \|^2 = \int_a^b (f(x) - P_m(x))^2 \, dx \ .$$

Sabemos, entretanto, que os polinômios de grau $\leq m$ constituem um espaço vetorial $K_m(x)$, do qual $\{1, x, x^2, \ldots, x^m\}$ é uma base. E mais: $K_m(x)$, para $x \in [a, b]$, é um sub-espaço de $C[a, b]$.

Se nos lembrarmos de projeção ortogonal de um vetor sobre um sub-espaço, (Capítulo 1), o nosso problema está resolvido, pois, a distância de f a P_m será mínima quando P_m for a projeção ortogonal de f sobre $K_m(x)$.

Resumindo: para aproximar $f \in C[a, b]$ por um polinômio $P_m(x)$ de grau no máximo m, basta determinar a projeção ortogonal de f sobre $K_m(x)$, o qual é gerado por $\{1, x, x^2, \ldots, x^m\}$.

Portanto os coeficientes $a_0, \ a_1, \ \ldots, \ a_m$, de $P_m(x)$, são dados pelo sistema normal, isto é:

$$\begin{pmatrix} (1,1) & (x,1) & \ldots & (x^m,1) \\ (1,x) & (x,x) & \ldots & (x^m,x) \\ \ldots\ldots & & & \\ (1,x^m) & (x,x^m) & \ldots & (x^m,x^m) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} (f,1) \\ (f,x) \\ \vdots \\ (f,x^m) \end{pmatrix} .$$

A menos que seja sugerido o produto escalar a ser utilizado, usa-se o produto escalar usual de $C[a, b]$, isto é, para $f, g \in C[a, b]$:

$$(f, g) = \int_a^b f(x) \ g(x) \ dx \ .$$

Exemplo 8.1 *Seja $f(x) = x^4 - 5\ x, \ x \in [-1, 1]$. Aproximar $f(x)$ por um polinômio do 2^o grau, usando o método dos mínimos quadrados.*

Solução: Temos que: $f(x) \in C[-1, 1]$, e para $x \in [-1, 1]$, $K_2(x)$ é um sub-espaço de $C[-1, 1]$. Queremos então:

$$f(x) \simeq P_2(x) = a_0 + a_1 x + a_2 x^2 \ ,$$

onde $P_2(x)$ deve ser a projeção ortogonal de f sobre $K_2(x)$. Assim a base para $K_2(x)$ é $\{1, x, x^2\}$. Devemos, então, resolver o sistema:

$$\begin{pmatrix} (1,1) & (x,1) & (x^2,1) \\ (1,x) & (x,x) & (x^2,x) \\ (1,x^2) & (x,x^2) & (x^2,x^2) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} (f,1) \\ (f,x) \\ (f,x^2) \end{pmatrix} .$$

Usando o produto escalar usual de $C[-1, 1]$, segue que:

$$
\begin{aligned}
(1,1) &= \int_{-1}^{1} dx = x]_{-1}^{1} = 2 , \\
(1,x) &= \int_{-1}^{1} x\ dx = \frac{x^2}{2}\Big]_{-1}^{1} = 0 = (x,1) , \\
\left(1,x^2\right) &= \int_{-1}^{1} x^2\ dx = \frac{x^3}{3}\Big]_{-1}^{1} = \frac{2}{3} = \left(x^2,1\right) = (x,x) , \\
\left(x,x^2\right) &= \int_{-1}^{1} x^3 dx = \frac{x^4}{4}\Big]_{-1}^{1} = 0 = \left(x^2,x\right) , \\
\left(x^2,x^2\right) &= \int_{-1}^{1} x^4 dx = \frac{x^5}{5}\Big]_{-1}^{1} = 2/5 , \\
(f,1) &= \int_{-1}^{1} (x^4 - 5\ x)\ dx = -\left(\frac{x^5}{5} - \frac{5\ x^2}{2}\right)\Big]_{-1}^{1} = -\frac{2}{5} , \\
(f,x) &= \int_{-1}^{1} (x^5 - 5\ x^2)\ dx = -\left(\frac{x^6}{6} - \frac{5\ x^3}{3}\right)\Big]_{-1}^{1} = -\frac{10}{3} , \\
\left((f,x^2\right) &= \int_{-1}^{1} (x^6 - 5\ x^3)\ dx = -\left(\frac{x^7}{7} - \frac{5\ x^4}{4}\right)\Big]_{-1}^{1} = \frac{2}{7} .
\end{aligned}
$$

Assim, obtemos:

$$
\begin{pmatrix} 2 & 0 & 2/3 \\ 0 & 2/3 & 0 \\ 2/3 & 0 & 2/5 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 2/5 \\ -10/3 \\ 2/7 \end{pmatrix} ,
$$

cuja solução é: $a_0 = -3/35$; $a_1 = -5$; $a_2 = 6/7$.

Portanto:

$$
f(x) \simeq P_2(x) = -\frac{3}{35} - 5\ x + \frac{6}{7}\ x^2. \tag{8.1}
$$

Exercícios

8.1 *Seja* $f(x) = \dfrac{1}{x+2}$, $x \in [-1,1]$. *Usando o método dos mínimos quadrados, aproximar a função* $f(x)$ *por um polinômio do* 2º *grau.*

8.2 *Seja* $f(x) = \dfrac{1}{x^4}$, $x \in [0,1]$. *Usando o método dos mínimos quadrados, aproximar a função* $f(x)$ *por um polinômio do tipo* $P(x) = a\ x^2 + b\ x^4$*, usando o seguinte produto escalar:*

$$
(f,g) = \int_0^1 x^2\ f(x)\ g(x)\ dx .
$$

Note que a base do sub-espaço neste caso é: $\{x^2, x^4\}$.

8.3 *Usando o método dos mínimos quadrados, aproximar a função* $f(x) = \left(x^3 - 1\right)^2$, $x \in [0,1]$,

a) *por uma reta;*

b) *por um polinômio do 2º grau.*

Observações:

a) Quem resolveu o último exercício (e quem não resolveu deve fazê-lo), pode observar que para passar de um polinômio de grau k para um polinômio de grau $k+1$ é necessário que calculemos todos os coeficientes do polinômio e não apenas o último, ou seja, devemos refazer praticamente todos os cálculos, visto que o sistema linear passa de ordem 2 para ordem 3.

b) Além disso, para m grande, $(m > 5)$, os efeitos de propagação dos erros de arredondamento, tornam-se explosivos, tornando o método tremendamente instável, ou seja, a solução do sistema linear pode ser irreal.

Vejamos então uma maneira de aumentar o grau do *polinômio aproximante* sem refazer todos os cálculos, bem como obter uma solução que realmente tenha significado.

Representação na Base Ortonormal

Veremos aqui como ajustar funções pelo método dos mínimos quadrados através de polinômios ortonormais.

Consideremos então em $K_m(x)$, uma base $\{L_0^*(x), L_1^*(x) \ldots, L_m^*(x)\}$ de polinômios ortonormais, isto é, de polinômios tais que:

$$\left(L_i^*(x), L_j^*(x)\right) = \delta_{ij} = \begin{cases} 1 & , \quad i = j \ , \\ 0 & , \quad i \neq j \ . \end{cases} \tag{8.2}$$

Observe que tais polinômios podem ser obtidos ortonormalizando-se a base canônica $\{1, x, x^2, \ldots, x^m\}$ por Gram-Schmidt, (Capítulo 1).

A projeção ortogonal de $f \in C[a,b]$ sobre $K_m(x)$ será então dada por:

$$P_m(x) = a_0 \ L_0^*(x) + a_1 \ L_1^*(x) + \ldots + a_m \ L_m^*(x),$$

onde os $a_i, \ i = 0, 1, \ldots, m$, são obtidos resolvendo-se o sistema:

$$\begin{pmatrix} (L_0^*, L_0^*) & (L_1^*, L_0^*) & \ldots & (L_m^*, L_0^*) \\ (L_0^*, L_1^*) & (L_1^*, L_1^*) & \ldots & (L_m^*, L_1^*) \\ \ldots\ldots & & & \\ (L_0^*, L_m^*) & (L_1^*, L_m^*) & \ldots & (L_m^*, L_m^*) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} (f, L_0^*) \\ (f, L_1^*) \\ \vdots \\ (f, L_m^*) \end{pmatrix} . \tag{8.3}$$

Mas, em vista de (8.2), (8.3) se reduz a:

$$\begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & & \ddots & \\ 0 & 0 & & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} (f, L_0^*) \\ (f, L_1^*) \\ \vdots \\ (f, L_m^*) \end{pmatrix} . \tag{8.4}$$

Agora a solução de (8.4) segue trivialmente, isto é:

$$a_i \ = \ (f, L_i^*) \quad , \quad i = 0, 1, \ldots, m. \tag{8.5}$$

Observe que se ao invés de uma base ortonormal, considerássemos uma base $L_i(x),\ i = 0, 1, \ldots, m$, **ortogonal**, cada $a_i,\ i = 0, 1, \ldots, m$, seria dado por:

$$a_i \ = \ \frac{(f, L_i)}{(L_i, L_i)} \quad .$$

Temos então obtido $P_m(x)$ que aproxima $f(x)$. Se agora quisermos aproximar $f(x)$ não só por $P_m(x)$ mas também por $P_{m+1}(x)$, devemos projetar $f(x)$ também sobre $K_{m+1}(x) \supset K_m(x)$.

Assim, uma base ortonormal para K_{m+1} será a base de $K_m(x)$ adicionada de $L^*_{m+1}(x)$ (ortonormal a $L_i^*(x),\ i = 0, 1, 2, \ldots, m$).

A projeção de $f(x)$ sobre $K_{m+1}(x)$ será:

$$P_{m+1} \ = \ a_0 \ L_0^*(x) \ + \ a_1 \ L_1^*(x) \ + \ \ldots \ + \ a_m \ L_m^*(x) \ + \ a_{m+1} \ L_{m+1}^*(x),$$

onde os a_i são dados por (8.5), inclusive a_{m+1}, isto é,

$$a_i \ = \ (f, L_i^*) \quad , \quad i = 0, 1, \ldots, m+1.$$

Observamos então que, uma vez obtido $P_m(x)$, basta calcularmos $L^*_{m+1}(x)$ e a_{m+1} para obter $P_{m+1}(x)$.

O processo pode ser repetido para $m+2, m+3, \ldots$.

Exemplo 8.2 - *Aproximar a função* $f(x) = x^4 - 5\ x,\ x \in [-1, 1]$,

a) *por uma reta,*

b) *por uma parábola,*

usando polinômios ortonormais.

Solução: A aproximação de $f(x)$ por uma reta será dada por:

$$f(x) \ \simeq \ a_0 \ P_0^*(x) \ + \ a_1 \ P_1^*(x) \ = \ Q_1(x) \ ,$$

e então a aproximação por uma parábola será obtida fazendo:

$$f(x) \ \simeq \ Q_1(x) \ + \ a_2 P_2^*(x) \ = \ Q_2(x) \ .$$

Devemos primeiramente construir os $P_i(x)$ (ortogonais) utilizando o processo de Gram-Schmidt a partir de $\{1, x, x^2\}$, (veja Capítulo 1).

Do **(Exemplo 1.14)**, temos que:

$$\begin{aligned} P_0(x) &= 1 \ , \\ P_1(x) &= x \ , \\ P_2(x) &= x^2 - \frac{1}{3} \ . \end{aligned}$$

Ortonormalizando primeiramente $P_0(x)$ e $P_1(x)$, para que possamos obter a reta que melhor aproxima $f(x)$, obtemos:

$$\begin{aligned} P_0^*(x) &= \frac{P_0}{[(P_0, P_0)]^{1/2}} = \frac{1}{\left[\int_{-1}^{1} dx\right]^{1/2}} = \frac{\sqrt{2}}{2} \ , \\ P_1^*(x) &= \frac{P_1}{[(P_1, P_1)]^{1/2}} = \frac{1}{\left[\int_{-1}^{1} x^2 \ dx\right]^{1/2}} = \frac{\sqrt{6}}{2} x \ . \end{aligned}$$

Assim os coeficientes a_0 e a_1 são dados por:

$$\begin{aligned} a_0 &= (f, P_0^*) = \int_{-1}^{1} \frac{\sqrt{2}}{2} (x^4 - 5\ x) dx = \frac{\sqrt{2}}{5} \ , \\ a_1 &= (f, P_1^*) = \int_{-1}^{1} \frac{\sqrt{6}}{2} \ x \ (x^4 - 5\ x) dx = - \frac{5\sqrt{6}}{3} \ . \end{aligned}$$

Portanto:

$$\begin{aligned} f(x) \simeq Q_1(x) &= \frac{\sqrt{2}}{5} \ P_0^*(x) - \frac{5\sqrt{3}}{3} \ P_1^*(x) \\ &= \frac{\sqrt{2}}{5} \left(\frac{\sqrt{2}}{2}\right) - \frac{5\sqrt{6}}{3} \left(\frac{\sqrt{6}}{2}\right) x \ . \end{aligned}$$

Devemos agora, ortonormalizar $P_2(x)$, (para obtermos a parábola). Assim:

$$\begin{aligned} P_2^*(x) &= \frac{P_2}{[(P_2, P_2)]^{1/2}} = \frac{(x^2 - 1/3)}{\left[\int_{-1}^{1} (x^2 - 1/3)^2 \ dx\right]^{1/2}} \\ &= \frac{3\sqrt{10}}{4} \ \left(x^2 - 1/3\right) \ . \end{aligned}$$

Então:

$$a_2 = (f, P_2^*) = \int_{-1}^{1} \frac{3\sqrt{10}}{4} \ \left(x^2 - 1/3\right) \ \left(x^4 - 5\ x\right) \ dx = \frac{4\sqrt{10}}{35} \ .$$

Portanto:

$$\begin{aligned} f(x) \;\simeq\; Q_2(x) &= Q_1(x) + \frac{4\sqrt{10}}{35} L_2^*(x) \\ &= \frac{\sqrt{2}}{5}\left(\frac{\sqrt{2}}{2}\right) - \frac{5\sqrt{6}}{3}\left(\frac{\sqrt{6}}{2}\right) x \\ &+ \frac{4\sqrt{10}}{35}\left[\frac{3\sqrt{10}}{4}\left(x^2 - 1/3\right)\right] . \end{aligned}$$

Observe que se agruparmos os termos semelhantes na última expressão obtemos exatamente (8.1), pois estaremos escrevendo a parábola em termos da base canônica de $K_2(x)$.

Exercícios

8.4 - *Aproximar, pelo método dos mínimos quadrados, a função* $f(x) = x^3 + 4\,x$, *no intervalo [0,1],*

a) *por uma reta,*

b) *por um polinômio do* 2º *grau,*

usando polinômios ortonormais.

8.5 - *Usando o método dos mínimos quadrados, aproximar a função* $f(x) = \left(x^2 - 3\,x + 4\right)^2$, $x \in [0,1]$,

a) *por uma reta;*

b) *por um polinômio do 2º grau,*

usando polinômios ortogonais.

8.2.2 Caso Discreto:

Vejamos agora o caso em que a função é dada por $n+1$ pares de pontos $(x_0, y_0),\ (x_1, y_1),\ \ldots,\ (x_n, y_n)$, onde $y_i = f(x_i),\ i = 0, 1, \ldots, n$, com os $n+1$ pontos $x_0, x_1, \ldots, x_n$ distintos.

Procuramos determinar um polinômio (a coeficientes reais)

$$P_m(x) = a_0 + a_1\,x + \ldots + a_m\,x^m , \tag{8.6}$$

de grau no máximo m , $(m < n)$, e tal que:

$$Q = \parallel f - P_m \parallel^2 ,$$

seja mínimo. Usando o produto escalar:

$$(f, g) \; = \; \sum_{k=0}^{n} \; f\,(x_k) \;\; g\,(x_k) \;\; ,$$

obtemos:

$$\begin{aligned} Q \;\; &= \;\; \| \; f - P_m \; \|^2 \, = \; (f - P_m, f - P_m) \\ &= \;\; \sum_{k=0}^{n} \left[f\,(x_k) - P_m\,(x_k) \right]^2 \; = \; \sum_{k=0}^{n} \; \left(y_k - P_m\,(x_k) \right)^2 \\ &= \;\; \sum_{k=0}^{n} \; \left(y_k - (a_0 + a_1 \; x_k + \ldots + a_m \; x_k^m) \right)^2 . \end{aligned}$$

Assim, dados os $n+1$ pontos distintos $x_0, \; x_1, \; \ldots, \; x_n$ e $n+1$ valores de uma função $y = f(x)$ sobre os pontos x_k, desejamos determinar um polinômio de grau no máximo m menor do que n tal que a soma dos quadrados dos desvios $y_k - P_m(x_k)$ entre os valores de $f(x)$ e $P_m(x)$ calculados nos pontos x_k, seja a menor possível.

Na verdade, precisamos determinar, na classe de todos os polinômios de grau $\leq m$, aquele que minimize Q.

O nosso problema resulta em última análise na determinação dos coeficientes $a_0, \; a_1, \; \ldots, \; a_m$ de $P_m(x)$.

Assim, coloquemos, por definição:

$$y \; = \; \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix} \; , \qquad\qquad p \; = \; \begin{pmatrix} P_m\,(x_0) \\ P_m\,(x_1) \\ \vdots \\ P_m\,(x_n) \end{pmatrix} \; ,$$

onde y e p são vetores do $I\!R^{n+1}$.

Usando (8.6), vemos que p pode ser escrito como:

$$p = a_0 \; \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \; + \; a_1 \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{pmatrix} \; + \; a_2 \; \begin{pmatrix} x_0^2 \\ x_1^2 \\ \vdots \\ x_n^2 \end{pmatrix} \; + \ldots + a_m \begin{pmatrix} x_0^m \\ x_1^m \\ \vdots \\ x_n^m \end{pmatrix} \; .$$

Denotando por:

$$u_0 \; = \; \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad ; \qquad u_i \; = \; \begin{pmatrix} x_0^i \\ x_1^i \\ \vdots \\ x_n^i \end{pmatrix} \qquad , \qquad i = 1, 2, \ldots, m \; ,$$

podemos escrever:

$$p \ = \ a_0 u_0 \ + \ a_1 u_1 \ + \ldots + \ a_m u_m \ .$$

Vamos mostrar agora que se os $n+1$ pontos são distintos, então os $m+1$ vetores $u_0, \ u_1, \ \ldots, \ u_m$ são linearmente independentes.

Para tanto, observe que p pode também ser escrito como:

$$p \ = \ \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ \ldots & \ldots & & & \\ 1 & x_m & x_m^2 & \ldots & x_m^m \\ \ldots & \ldots & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{pmatrix} \quad \begin{pmatrix} a_0 \\ a_1 \\ \\ \vdots \\ a_m \end{pmatrix} .$$

Seja A a matriz dos coeficientes; isto é:

$$A \ = \ \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ \ldots & \ldots & & & \\ 1 & x_m & x_m^2 & \ldots & x_m^m \\ \ldots & \ldots & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{pmatrix} .$$

A matriz A possui $n+1$ linhas por $m+1$ colunas, com $n > m$. Seja A', a submatriz quadrada constituída das $m+1$ primeiras linhas e $m+1$ primeiras colunas de A. Assim:

$$A' \ = \ \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ \ldots & \ldots & & & \\ 1 & x_m & x_m^2 & \ldots & x_m^m \end{pmatrix} .$$

A matriz A' é tal que $det \ A' = \prod_{i>j} (x_i - x_j)$. Desde que os pontos $x_0, \ x_1, \ \ldots, \ x_n$ são distintos segue que $detA' \neq 0$. (Este fato está demonstrado no Capítulo 10).Então existe uma submatriz de A, de ordem $m+1$, que é não singular. Assim, os vetores $u_0, \ u_1, \ \ldots, \ u_m$ são linearmente independentes.

Portanto $u_0, u_1, \ldots, u_m$ geram em $I\!R^{n+1}$ um sub-espaço vetorial V de dimensão $m+1 < n+1$ (pois $m < n$, por hipótese).

Temos que: $y \in I\!R^{n+1}$ e $p \in V \subset I\!R^{n+1}$ e queremos que a distância de y a p seja mínima. Isto ocorrerá quando p for a projeção ortogonal de y sobre V.

Os coeficientes $a_0, a_1, \ldots, a_m$ do polinômio procurado são então dados pelo sistema normal:

$$\begin{pmatrix} (u_0, u_0) & (u_1, u_0) & \ldots & (u_m, u_0) \\ (u_0, u_1) & (u_1, u_1) & \ldots & (u_m, u_1) \\ \ldots & \ldots & & \\ (u_0, u_m) & (u_1, u_m) & \ldots & (u_m, u_m) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} (y, u_0) \\ (y, u_1) \\ \vdots \\ (y, u_m) \end{pmatrix} .$$

A menos que seja sugerido o produto escalar a ser utilizado, usa-se o produto escalar usual do $I\!R^{n+1}$, isto é:

$$(x, y) \; = \; \sum_{i=1}^{n} x_i \; y_i,$$

onde $x = (x_0, \; x_1, \; \ldots, \; x_n)^t$ e $y = (y_0, \; y_1, \; \ldots, \; y_n)^t$.

Exemplo 8.3 *Dada a função $y = f(x)$, por meio da tabela:*

x	-1	0	1	2
y	0	-1	0	7

ajustá-la por um polinômio do 2º grau, usando o método dos mínimos quadrados.

Solução: Neste caso queremos: $f(x) \;\; \simeq \;\; P_2(x) \; = \; a_0 + a_1 \; x + a_2 \; x^2$. Assim, devemos construir $p \; = \; a_0 \; u_0 + a_1 \; u_1 + a_2 \; u_2$.

Fazendo:

$$y = \begin{pmatrix} 0 \\ -1 \\ 0 \\ 7 \end{pmatrix} \; ; \; u_0 \; = \; \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \; ; \; u_1 \; = \; \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \end{pmatrix} \; ; \; u_2 \; = \; \begin{pmatrix} 1 \\ 0 \\ 1 \\ 4 \end{pmatrix} \; ,$$

devemos resolver o sistema:

$$\begin{pmatrix} (u_0, u_0) & (u_1, u_0) & (u_2, u_0) \\ (u_0, u_1) & (u_1, u_1) & (u_2, u_1) \\ (u_0, u_2) & (u_1, u_2) & (u_2, u_2) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} (y, u_0) \\ (y, u_1) \\ (y, u_2) \end{pmatrix} .$$

Usando o produto escalar usual do $I\!R^4$, segue que:

$$\begin{aligned}
(u_0, u_0) &= 1 + 1 + 1 + 1 \; = \; 4 \; , \\
(u_0, u_1) &= -1 + 0 + 1 + 2 \; = \; 2 \; = \; (u_1, u_0) \; , \\
(u_0, u_2) &= 1 + 0 + 1 + 4 \; = \; 6 \; = \; (u_2, u_0) \; , \\
(u_1, u_1) &= 1 + 0 + 1 + 4 \; = \; 6 \; , \\
(u_1, u_2) &= -1 + 0 + 1 + 8 \; = \; 8 \; = \; (u_2, u_1) \; , \\
(u_2, u_2) &= 1 + 0 + 1 + 16 \; = \; 18 \; , \\
(y, u_0) &= 0 + -1 + 0 + 7 \; = \; 6 \; , \\
(y, u_1) &= 0 + 0 + 0 + 14 \; = \; 14 \; , \\
(y, u_2) &= 0 + 0 + 0 + 28 \; = \; 28 \; .
\end{aligned}$$

Obtemos então o sistema linear:

$$\begin{pmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 14 \\ 28 \end{pmatrix} ,$$

cuja solução é: $a_0 = -\frac{8}{5}$; $a_1 = \frac{1}{5}$; $a_2 = 2$.

Portanto a parábola que melhor aproxima a função tabelada é:

$$P_2(x) = -\frac{8}{5} + \frac{1}{5}\, x + 2\, x^2 .$$

Exercícios

8.6 - *Determinar, pelo método dos mínimos quadrados, a reta mais próxima dos pontos (x_i, y_i) para a função $y = f(x)$ tabelada:*

x	-2	-1	0	1	2
y	0	0	-1	0	7

8.7 - *Determinar a parábola mais próxima dos pontos (x_i, y_i) para a função $y = f(x)$ tabelada:*

x	-3	-1	1	2	3
y	-1	0	1	1	-1

usando o método dos mínimos quadrados.

8.8 - *Usando o método dos mínimos quadrados, aproxime a função dada pela tabela:*

x	0	1	2	3	4	5
y	-1	0	3	8	15	24

por um polinômio do tipo:$P(x) = a + b\, x^3$, *usando o produto escalar:*

$$(x, y) = \sum_{i=0}^{n} (i+1)\, x_i\, y_i .$$

8.2.3 Erro de Truncamento

O erro de truncamento no método dos mínimos quadrados é dado por **Q**.

Assim temos:

a) caso contínuo:

$$Q = \| f - P_m \|^2 = \int_a^b (f(x) - P_m(x))^2\, dx .$$

b) caso discreto:

$$Q = \| f - P_m \|^2 = \sum_{k=0}^{n} \left(y_k - P_m \left(x_k \right) \right)^2 \ .$$

Para ilustrar, calculemos o erro de truncamento no último exemplo. Assim:

$$\begin{aligned} Q &= \sum_{k=0}^{3} \left(y_k - P_2 \left(x_k \right) \right)^2 \\ &= \sum_{k=0}^{3} y_k^2 - 2 \sum_{k=0}^{3} y_k P_2 \left(x_k \right) + \sum_{k=0}^{3} P_2^2 \left(x_k \right) \\ &= 50 - \frac{492}{5} + \frac{1230}{25} = \frac{4}{5} = 0,8 \ . \end{aligned}$$

Observações:

a) O valor encontrado ($Q = 0.8$) corresponde à soma dos quadrados dos desvios entre os valores da função e do polinômio calculados nos pontos tabelados. Além disso, podemos afirmar que a parábola encontrada é a *melhor* entre as equações do 2º grau, ou seja, para qualquer outra parábola teremos para Q um valor maior do que o encontrado.

b) Em muitos casos, os dados experimentais não se assemelham a polinômios. Se faz necessário então procurar funções (não polinomiais) que melhor aproximem os dados. Trataremos nas próximas duas seções de como aproximar uma função dada por funções que não sejam polinômios.

8.3 Aproximação Trigonométrica

Suponha que a função a ser aproximada tem caracteristícas de periodicidade. Neste caso, aproximar por polinômio não seria uma boa escolha, visto que estes não possuem a propriedade de serem periódicos. Devemos portanto tentar aproximar a função por uma outra que tenha propriedades semelhantes à função que estamos querendo aproximar.

8.3.1 Caso Contínuo

Consideremos uma função $f(x)$ periódica e integrável em $[0, 2\pi]$.

Desejamos aproximar $f(x), \ x \in [0, 2\pi]$, por uma função do tipo:

$$\begin{aligned} f(x) &\simeq a_0 + a_1 \cos x + b_1 \ sen \ x + a_2 \cos 2x + b_2 \ sen \ 2x \\ &+ \ldots + a_m \cos mx + b_m \ sen \ mx = F(x) \ . \end{aligned} \tag{8.7}$$

de tal modo que a distância de f a F seja mínima. A aproximação da forma (8.7) é chamada **aproximação trigonométrica** de ordem m, para $f(x)$.

Adotando o produto escalar:

$$(f, g) = \int_0^{2\pi} f(x) \ g(x) \ dx \ , \tag{8.8}$$

nosso objetivo é minimizar:

$$Q \ = \parallel f - F \parallel^2 = \int_0^{2\pi} (f(x) - F(x))^2 \ dx \ .$$

Evidentemente, tal objetivo será alcançado quando F for a projeção ortogonal de f sobre o subespaço gerado por:

$$\{1, \cos\ x, sen\ x, \cos\ 2x, sen\ 2x, \ldots, \cos\ mx, sen\ mx\} \ . \tag{8.9}$$

Assim, os coeficientes: $a_0, a_1, b_1, \ldots, a_m, b_m$ são determinados resolvendo-se o sistema normal:

$$\begin{pmatrix} (1,1) & (\cos\ x, 1) & \ldots & (sen\ mx, 1) \\ (1, \cos\ x) & (\cos\ x, \cos\ x) & \ldots & (sen\ mx, \cos\ x) \\ \ldots & \ldots & & \\ (1, sen\ mx) & (\cos\ x, sen\ mx) & \ldots & (sen\ mx, sen\ mx) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ b_m \end{pmatrix}$$

$$= \begin{pmatrix} (f,1) \\ (f, \cos\ x) \\ \vdots \\ (f, sen\ mx) \end{pmatrix} \ . \tag{8.10}$$

A sequência (8.9) é ortogonal em $[0, 2\pi]$, isto é:

$$\int_0^{2\pi} sen\ px\ \ \cos\ qx\ dx \ = \ 0 \ ,$$

e temos:

$$\int_0^{2\pi} sen\ px\ sen\ qx\ dx \ = \begin{cases} 0 & p \neq q \ , \\ \pi & p = q \neq 0 \ , \end{cases}$$

$$\int_0^{2\pi} \cos px\ \ \cos qx\ dx \ = \begin{cases} 0 & p \neq q \ , \\ \pi & p = q \neq 0 \ , \\ 2\pi & p = q = 0 \ . \end{cases}$$

Portanto o sistema (8.10), se reduz a:

$$\begin{pmatrix} 2\pi & & & \\ & \pi & & \bigcirc \\ & & \ddots & \\ \bigcirc & & & \pi \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} (f,1) \\ (f, \cos\ x) \\ \vdots \\ (f, sen\ mx) \end{pmatrix} \ ,$$

cuja solução é dada por:

$$\begin{aligned} a_0 &= \frac{1}{2\pi}\int_0^{2\pi} f(x)\ dx\ , \\ a_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\ \cos\ kx\ dx\ ,\ k = 1, 2, \ldots, m\ , \\ b_k &= \frac{1}{\pi}\int_0^{2\pi} f(x)\ sen\ kx\ dx\ ,\ k = 1, 2, \ldots, m\ . \end{aligned}$$

Observações:

i) Se a função dada é uma função par, então, desde que *sen* x é uma função impar, temos que:

$$\int_0^{2\pi} f(x)\ sen\ x\ dx\ =\ 0\ ,$$

e portanto:

$$f(x)\ \simeq\ a_0 + \sum_{k=1}^{m} a_k\ \cos\ kx\ .$$

ii) Se a função dada é uma função impar, então, desde que $\cos\ x$ é uma função par, temos que:

$$\int_0^{2\pi} f(x)\ \cos\ x\ dx\ =\ 0\ ,$$

e portanto:

$$f(x)\ \simeq\ \sum_{k=1}^{m} b_k\ sen\ kx\ .$$

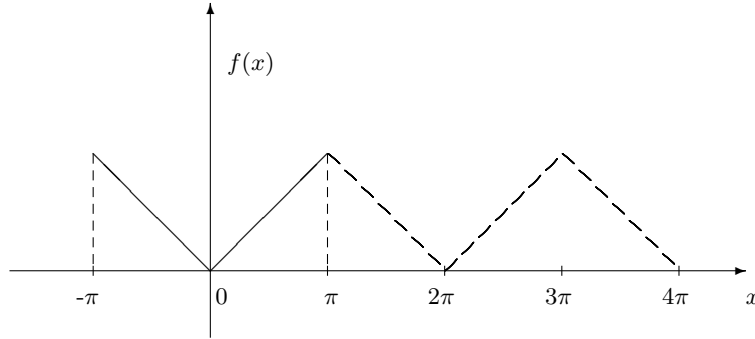
iii) A aproximação de $f(x)$ no intervalo $[0, 2\pi]$ por (8.7), usando o produto escalar (8.8), pelo método dos mínimos quadrados é também conhecido como **Análise Harmônica**. Os termos: $a_k\ \cos\ kx + b_k\ sen\ kx$ podem ser expressos na forma: $A_k\ sen(kx + \psi_k)$, onde

$$A_k\ =\ \sqrt{(a_k^2 + b_k^2)}\ \text{ e }\ tg\ \psi_k\ =\ \frac{a_k}{b_k}\ ,$$

e são chamados *harmônicos de ordem k*. O termo A_k é denominado *amplitude* e ψ_k *ângulo de fase*.

Exemplo 8.4 - *Obter aproximação trigonométrica de ordem 1, para:*

$$f(x)\ =\ |x|,\ -\pi \le x \le \pi\ .$$

Solução: Prolongando a função dada, periodicamente, (período 2π), obtemos:

Devemos determinar então $F(x)$ tal que:

$$f(x) \simeq a_0 + a_1 \cos x + b_1 \, sen \, x = F(x) \; .$$

Observe que integrar de $-\pi$ a π é igual a integrar de 0 a 2π e que a função dada é uma função par, logo $b_1 = 0$ e portanto:

$$f(x) \simeq a_0 + a_1 \cos x = F(x) \; .$$

Assim:

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_0^{2\pi} f(x) \, dx = \frac{2}{2\pi} \int_0^{\pi} x \, dx = \frac{\pi}{2} \\ a_1 &= \frac{1}{\pi} \int_0^{2\pi} x \cos x \, dx = \frac{2}{\pi} \int_0^{\pi} x \cos x \, dx \\ &= \frac{2}{\pi} \left[x \, sen \, x \Big]_0^{\pi} - \int_0^{\pi} sen \, x \, dx \right] \\ &= 0 + \frac{2}{\pi} \cos x \Big]_0^{\pi} = -\frac{4}{\pi} \; . \end{aligned}$$

Portanto, obtemos que a aproximação trigonométrica de ordem 1 para a função dada é:

$$f(x) \simeq \frac{\pi}{2} - \frac{4}{\pi} \cos x \; .$$

Exercícios:

8.9 - *Mostre que o conjunto das funções:*

$$\{1, \quad \cos t, \quad sen \; t, \quad \cos 2t, \quad sen \; 2t, \ldots, \quad sen \; mt\} \; ,$$

é um sistema ortogonal em $[-\pi, \pi]$.

8.10 - *Considere a função:*

$$y(t) = \begin{cases} -1 \ , & -\pi \leq t \leq 0 \ ; \\ 1 \ , & 0 \leq t \leq \pi \ . \end{cases}$$

Verificar que a aproximação trigonométrica de grau 2:

$$y(t) = a_0 + a_1 \cos t + b_1 \; sen \; t + a_2 \cos 2t + b_2 \; sen \; 2t$$

para y(t) é dada por: $y(t) = \frac{4}{\pi} sen 2t$.

Obs: Estenda y(t) a uma função impar de período 2π.

8.3.2 Caso Discreto

Consideremos uma função $f(x)$ conhecida nos N pontos distintos:

$$x_k = \frac{2k\pi}{N} \ , \ k = 1, 2, \ldots, N \ .$$

Desejamos aproximar $f(x)$, por uma função do tipo:

$$\begin{aligned} f(x) &\simeq a_0 + a_1 \cos x + b_1 \; sen \; x + a_2 \cos 2x + b_2 \; sen \; 2x \\ &+ \ldots + a_L \cos Lx + b_L \; sen \; Lx = S_L(x) \ . \end{aligned} \tag{8.11}$$

com $L \leq \frac{N}{2}$; de tal modo que a distância de f a S_L seja mínima.

Adotando o produto escalar:

$$(f, g) = \sum_{k=1}^{N} f(x_k) \; g(x_k) \ . \tag{8.12}$$

nosso objetivo é minimizar:

$$Q = \parallel f - S_L \parallel^2 = \sum_{k=1}^{N} (f(x_k) - S_L(x_k))^2 \ .$$

Evidentemente, tal objetivo será alcançado quando S_L for a projeção ortogonal de f sobre o subespaço gerado por:

$$\{1, \ \cos x, \ sen \; x, \ \cos 2x, \ sen \; 2x, \ \ldots, \ \cos Lx, \ sen \; Lx\} \ . \tag{8.13}$$

Assim, os coeficientes: $a_0, a_1, b_1, \ldots, a_L, b_L$ são determinados resolvendo-se o sistema (8.10), onde o produto escalar é dado por (8.12).

A sequência (8.13) é ortogonal em $[0, 2\pi]$, isto é:

$$\sum_{k=1}^{N} sen \; px_k \; \cos \; qx_k = 0, \tag{8.14}$$

e temos:

$$\sum_{k=1}^{N} sen\ px_k\ sen\ qx_k = \begin{cases} 0 & p \neq q\ , \\ \frac{N}{2} & p = q \neq 0. \end{cases}$$

$$\sum_{k=1}^{N} \cos\ px_k\ \cos\ qx_k = \begin{cases} 0 & p \neq q\ , \\ N & p = q \neq 0\ , \\ \frac{N}{2} & p = q = 0\ . \end{cases}$$

Observe que o sistema normal aqui é o mesmo do caso contínuo, o que muda é o produto escalar, pois nesse caso temos a função tabelada. Portanto o sistema normal se reduz a:

$$\begin{pmatrix} N & & & \bigcirc \\ & \frac{N}{2} & & \\ & & \ddots & \\ \bigcirc & & & \frac{N}{2} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} (f,1) \\ (f, \cos\ x) \\ \vdots \\ (f, sen\ Lx) \end{pmatrix}\ ,$$

cuja solução é dada por:

$$a_0 = \frac{1}{N} \sum_{k=1}^{N} f(x_k)\ ,$$

$$a_j = \frac{2}{N} \sum_{k=1}^{N} f(x_k)\ \cos\ jx_k\ ,\ k = 1, 2, \ldots, L\ ,$$

$$b_j = \frac{2}{N} \sum_{k=1}^{N} f(x_k)\ sen\ jx_k\ ,\ k = 1, 2, \ldots, L\ .$$

Exemplo 8.5 - *Obter aproximação trigonométrica de ordem 1, para a função dada pela tabela :*

x	$\frac{\pi}{4}$	$\frac{\pi}{2}$	$\frac{3\pi}{4}$	π	$\frac{5\pi}{4}$	$\frac{3\pi}{2}$	$\frac{7\pi}{4}$	2π
$f(x)$	126	159	191	178	183	179	176	149

usando o método dos mínimos quadrados.

Solução: Note que os pontos tabelados são: $x_k = \frac{2k\pi}{8},\ k = 1, 2, \ldots, 8.$

Devemos determinar então $S_L(x)$ tal que:

$$f(x) \simeq a_0 + a_1\ \cos\ x + b_1\ sen\ x = S_1(x)\ .$$

Assim:

$$a_0 = \frac{1}{8} \sum_{k=1}^{8} f(x_k) = 167.625$$

$$a_1 = \frac{2}{8} \sum_{k=1}^{8} f(x_k) \cos x_k = -19.978$$

$$b_1 = \frac{2}{8} \sum_{k=1}^{8} f(x_k) \; sen \; x_k = -12.425$$

Portanto, obtemos que a aproximação trigonométrica de ordem 1 para a função dada é:

$$f(x) \simeq 167.625 - 19.978 \cos x - 12.425 \; sen \; x \, .$$

Exercício

8.11 - *Considere a função $f(x)$, dada por:*

k	1	2	3	4	5	6
$f(x)$	11.8	4.3	13.8	3.9	−18.1	−22.9
	7	8	9	10	11	12
	−27.2	−23.8	8.2	31.7	34.2	38.4

onde $x_k = \frac{2k\pi}{12}$. *Obtenha aproximação trigonométrica, de ordem 2, usando o método dos mínimos quadrados.*

Observe que se utilizarmos o produto escalar dado por (8.12) com $x_k = \frac{2k\pi}{N}, \; k = 1, 2, \ldots, N$, estaremos fazendo uma análise harmônica aproximada.

8.4 Outros Tipos de Aproximação

O objetivo dos métodos dos mínimos quadrados é aproximar a função dada por uma família *linear nos parâmetros*, isto é, definidas por expressões da forma:

$$a_0 \, g_0(x) + a_1 \, g_1(x) + \ldots + a_m \, g_m(x) \, .$$

Muitos casos podem ser reduzidos a essa forma por uma transformação prévia do problema, isto é, nosso objetivo agora consiste na linearização do problema, através de transformações convenientes, de modo que o método dos mínimos quadrados possa ser aplicado. Daremos a seguir alguns exemplos que ilustram o problema a ser resolvido.

1^o **caso:** Aproximar $f(x)$ por uma função do tipo exponencial, isto é,

$$f(x) \simeq a \, b^x \, ,$$

pode ser reduzido por uma transformação logarítmica ao de aproximar:

$$ln\ f(x) \simeq ln\ a + x\ ln\ b\ .$$

Fazendo:

$$F(x) = ln\ f(x);$$

$$a_0 = \ln\ a; \quad a_1 = ln\ b;$$

$$g_0(x) = 1; \quad g_1(x) = x,$$

reduzimos o problema original ao de aproximar $F(x)$ por $a_0\ g_0(x) + a_1\ g_1(x)$.

Assim a_0 e a_1 serão obtidos resolvendo-se o sistema linear:

$$\begin{pmatrix} (g_0(x), g_0(x)) & (g_1(x), g_0(x)) \\ (g_0(x), g_1(x)) & (g_1(x), g_1(x)) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} (F(x), g_0(x)) \\ (F(x), g_1(x)) \end{pmatrix}. \tag{8.15}$$

Obtidos os valores de a_0 e a_1, a aplicação da função exponencial em $a_0 = \ln\ a$ e $a_1 = ln\ b$ retorna os valores originais requeridos, isto é: $a = e^{a_0}$ e $b = e^{a_1}$.

Observe que em todos os casos desta seção, o método dos mínimos quadrados será então aplicado ao problema linearizado, seguido de um retorno aos parâmetros originais. Portanto os parâmetros assim obtidos não serão, em geral, *ótimos* no sentido do método dos mínimos quadrados, pois este não terá sido aplicado ao problema original.

2º **caso:** Aproximar $f(x)$ por uma função do tipo geométrica, isto é,

$$f(x) \simeq a\ x^b\ ,$$

pode ser reduzido por uma transformação logarítmica ao de aproximar

$$ln\ f(x) \simeq ln\ a + b\ ln\ x\ .$$

Fazendo:

$$F(x) = ln\ f(x);$$

$$a_0 = \ln\ a; \quad a_1 = b;$$

$$g_0(x) = 1; \quad g_1(x) = ln\ x,$$

reduzimos o problema original ao de aproximar $F(x)$ por $a_0\ g_0(x) + a_1\ g_1(x)$. Assim a_0 e a_1 serão obtidos resolvendo-se o sistema linear (8.15). Nesse caso a volta aos parâmetros originais será feita usando exponencial para determinação de a, e b será simplesmente o valor de a_1.

3º **caso:** Aproximar $f(x)$ por uma função do tipo hiperbólica, isto é,

$$f(x) \simeq \frac{1}{a + b\,x},$$

pode ser reduzido ao problema de aproximar:

$$\frac{1}{f(x)} \simeq a + b\,x.$$

Fazendo:

$$F(x) = \frac{1}{f(x)};$$

$$a_0 = a; \quad a_1 = b;$$

$$g_0(x) = 1; \quad g_1(x) = x,$$

reduzimos o problema original ao de aproximar $F(x)$ por $a_0\ g_0(x) + a_1\ g_1(x)$. Assim a_0 e a_1 serão obtidos resolvendo-se o sistema linear (8.15). Nesse caso os parâmetros originais serão simplesmente o valor de a_0 e de a_1.

Já deu para notar como deve ser resolvido o problema. Assim nos próximos casos apenas faremos a parte da transformação a ser realizada.

4º **caso:** Aproximar $f(x)$ por uma função do tipo $\sqrt{a + b\,x}$, isto é:

$$f(x) \simeq \sqrt{a + b\,x},$$

pode ser reduzido ao problema de aproximar:

$$f^2(x) \simeq a + b\,x.$$

Fazemos então:

$$F(x) = f^2(x);$$

$$a_0 = a; \quad a_1 = b;$$

$$g_0(x) = 1; \quad g_1(x) = x,$$

5º **caso:** Aproximar $f(x)$ por uma função do tipo $x\ ln\ (a + b\,x)$, isto é,

$$f(x) \simeq x\ ln\ (a + b\,x),$$

pode ser reduzido ao problema de aproximar:

$$e^{\frac{f(x)}{x}} \simeq a + b\,x.$$

Fazemos então:

$$F(x) = e^{\frac{f(x)}{x}};$$

$$a_0 = a; \quad a_1 = b;$$

$$g_0(x) = 1; \quad g_1(x) = x \ .$$

Daremos a seguir alguns exemplos.

Exemplo 8.6 *Aproximar a função tabelada:*

x	0	1	2	3
y	1	1	1.7	2.5

por uma função racional, isto é:

$$f(x) \simeq \frac{a + x^2}{b + x} \ .$$

Solução: Em primeiro lugar devemos rearranjar a equação de tal forma que o lado direito seja combinação linear de funções conhecidas. Assim:

$$y(x) \simeq \frac{a + x^2}{b + x}$$

$$\Rightarrow (b + x)\ y(x) \simeq a + x^2$$

$$\Rightarrow b\ y(x) \simeq a - x\ y(x) + x^2$$

$$\Rightarrow y(x) \simeq \frac{a}{b} - \frac{1}{b}[x\ (y(x) - x)]$$

Fazendo:

$$F(x) = y(x);$$

$$a_0 = \frac{a}{b}; \quad a_1 = -\frac{1}{b};$$

$$g_0(x) = 1; \quad g_1 = x(y(x) - x),$$

obtemos que $F(x) \simeq a_0\ g_0(x) + a_1\ g_1(x)$. Assim, devemos resolver o sistema (8.15), onde:

$$F(x) = \begin{pmatrix} 1 \\ 1 \\ 1.7 \\ 2.5 \end{pmatrix} \ ; \ g_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \ ; \ g_1 = \begin{pmatrix} 0 \\ 0 \\ -0.6 \\ -1.5 \end{pmatrix} \ .$$

Usando o produto escalar usual, obtemos:

$$\begin{pmatrix} 4 & -2.1 \\ -2.1 & 2.61 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 6.2 \\ -4.77 \end{pmatrix} ,$$

cuja solução é: $a_0 = 1.0224$ e $a_1 = -1.005$.

Agora, desde que:

$$a_1 = -\frac{1}{b} \Rightarrow b = -\frac{1}{a_1}$$

$$\Rightarrow b = 0.9950 ,$$

$$a_0 = \frac{a}{b} \Rightarrow a = a_0 \times b$$

$$\Rightarrow a = 1.0173 ,$$

obtemos que:

$$F(x) \simeq 1.0224 - 1.005[x(y(x) - x)] ,$$

$$y(x) \simeq \frac{1.0173 + x^2}{0.9950 + x} .$$

Note que a função minimizada foi:

$$Q = \sum_{i=0}^{3} [y_i - (1.0224 - 1.005\ x_i(y_i - x_i))]^2 ,$$

isto é, foi minimizado o quadrado da diferença entre a função $F(x) = y(x)$ e a função $a_0 + a_1[x(y(x) - x)]$.

Exemplo 8.7 - *A intensidade de uma fonte radioativa é dada por:*

$$I = I_0\ e^{-\alpha t} .$$

Através de observações, tem-se:

t	0.2	0.3	0.4	0.5	0.6	0.7	0.8
I	3.16	2.38	1.75	1.34	1.00	0.74	0.56

Determinar I_0 e α.

Solução: Novamente devemos rearranjar a equação. Assim:

$$I \simeq I_0\ e^{-\alpha t}$$

$$\Rightarrow \quad ln\ I \simeq ln\ I_0 - \alpha\ t .$$

Fazendo, então:

$$F(t) = ln\ I;$$

$$a_0 = ln\ I_0; \quad a_1 = -\alpha;$$

$$g_0(t) = 1; \quad g_1(t) = t,$$

obtemos que $F(t) \simeq a_0\, g_0(t) + a_1\, g_1(t)$. Assim, devemos resolver o sistema (8.15), onde:

$$F(t) = \begin{pmatrix} 1.15 \\ 0.87 \\ 0.56 \\ 0.30 \\ 0.00 \\ -0.30 \\ -0.58 \end{pmatrix} ; \; g_0(t) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} ; \; g_1(t) = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \\ 0.6 \\ 0.7 \\ 0.8 \end{pmatrix} .$$

Usando o produto escalar usual, obtemos:

$$\begin{pmatrix} 7 & 3.5 \\ 3.5 & 2.03 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 2.0 \\ 0.19 \end{pmatrix} .$$

cuja solução é: $a_0 = 1.7304$ e $a_1 = -2.8893$.

Agora, desde que:

$$a_1 = -\alpha \Rightarrow \alpha = 2.8893 ,$$

$$a_0 = ln\ I_0 \Rightarrow I_0 = e^{a_0}$$

$$\Rightarrow I_0 = 5.6429 ,$$

obtemos:

$$F(t) \simeq 1.7304 - 2.8893\, t ,$$

$$I \simeq 5.6429\, e^{-2.8893\, t} .$$

A função minimizada foi:

$$Q = \sum_{i=0}^{6} \{F(t_i) - (1.73 - 2.89 t_i)\}^2 .$$

Observe que para os casos onde a função f é determinada empiricamente por pontos observados, o problema que surge é: qual família ajusta melhor os dados?

A seleção de funções pode ser feito percorrendo-se as seguintes etapas:

i) Selecionar a priori um número pequeno de famílias de funções. (A escolha das famílias deve ser tal que os membros sejam de fácil avaliação para argumentos dados).

ii) Utilizar todas as informações disponíveis sobre f para eliminar as famílias que obviamente não são adequadas.

iii) Aplicar às famílias que restarem o chamado *teste de alinhamento*. (Este teste está descrito a seguir).

iv) Determinar os parâmetros para as famílias que passaram pelo teste de alinhamento e escolher o melhor resultado por comparação.

Teste de Alinhamento:

O teste de alinhamento consiste em:

i) Transformar a equação da família $y = f(x)$ em, por exemplo, outras duas famílias: $F_1(x) = a_0\ g_0(x) + a_1\ g_1(x)$ onde $g_0(x) = 1$ e a_0 e a_1 não dependem de $F_1(x)$ e $g_1(x)$ e $F_2(x) = a_0'\ g_0'(x) + a_1'\ g_1'(x)$ onde $g_0'(x) = 1$ e a_0' a_1' não dependem de $F_2(x)$ e $g_1'(x)$.

ii) Fazer os gráficos: (I) $g_1(x)$ contra $F_1(x)$ e (II) $g_1'(x)$ contra $F_2(x)$.

iii) Escolher a família $F_1(x)$ se o gráfico (I) estiver mais linear que o gráfico (II), e $F_2(x)$ se o gráfico (II) for mais linear que (I).

Observe que devido aos erros de observação que afetam x_k e $f(x_k)$, não excluiremos a família se os pontos no gráfico se distribuirem aleatoriamente em torno de uma reta média.

Exemplo 8.8 - *Considere a função dada pela tabela:*

t	-8	-6	-4	-2	0	2	4
y	30	10	9	6	5	4	4

Qual das funções:

$$\textbf{a)}\quad y(t) = \frac{1}{a+bt}\ ; \qquad \textbf{b)}\quad y(t) = ab^t\ ,$$

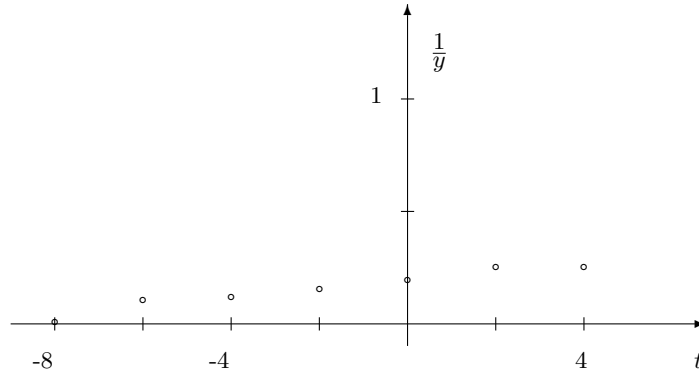
ajustaria melhor os dados da tabela?

Solução: Em primeiro lugar devemos linearizar as funções **a)** e **b)**. Assim de $y(t) = \frac{1}{a+bt}$, obtemos:

$$\textbf{(I)}\quad \frac{1}{y(t)} \simeq a + b\ t\ .$$

Fazendo:

$$\begin{aligned} F_1(t) &= \frac{1}{y(t)}; \\ a_0 &= a;\ \ a_1 \ = \ b; \\ g_0(t) &= 1;\ \ g_1(t) \ = \ t, \end{aligned}$$

obtemos que $F_1(t) \simeq a_0\ g_0(t) + a_1\ g_1(t)$. Assim devemos fazer o gráfico de t contra $\frac{1}{y}$:

Para $y(t) = a\ b^t$ obtemos:

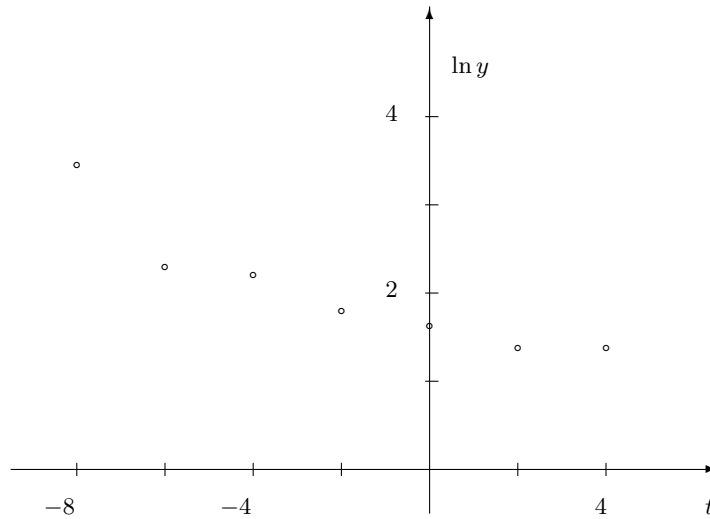
$$(II) \quad ln\ y(t) \simeq ln\ a + t\ ln\ b\ .$$

Fazendo:

$$F_2(t) \ = \ ln\ y(t);$$

$$a_0 \ = \ ln\ a; \quad a_1 \ = \ ln\ b;$$

$$g_0'(t) \ = \ 1; \quad g_1'(t) \ = \ t,$$

obtemos que $F_2(t) \ \simeq \ a_0\ g_0'(t) \ + \ a_1\ g_1'(t)$. Assim devemos fazer o gráfico de t contra $ln\ y(t)$:

Vemos que o gráfico de t contra $\dfrac{1}{y(t)}$ é mais linear que o gráfico de t contra $ln\ y(t)$. Assim, devemos escolher $y = \dfrac{1}{a + b\ t}$ para aproximar a função dada.

Note que o problema só estará resolvido se pudermos linearizar a função dada. Entretanto, existem casos onde a função não pode ser linearizada diretamente, como mostra o exemplo a seguir.

Exemplo 8.9 - *Aproximar $f(x)$ dada pela tabela:*

x	0	2	4	6	8	10	∞
$f(x)$	84.8	75.0	67.2	61.9	57.6	53.4	20.0

por função do tipo: $\dfrac{1}{a+b\,x}+c$.

Solução: Não existe uma transformação que torne o lado direito (da aproximação) combinação linear de funções conhecidas. Podemos entretanto, tentar resolver o problema analisando a função f. Neste caso, sobre a função f, conhecemos os valores dados na tabela e f tem como limite para $x \to \infty$ o valor 20. Como

$$\lim_{x\to\infty} \frac{1}{a+b\,x} + c \; = \; c \;,$$

adotaremos $c = 20$. Portanto, podemos escrever:

$$f(x) \simeq \frac{1}{a+b\,x} + 20$$

$$\Rightarrow \quad f(x) - 20 \simeq \frac{1}{a+b\,x}$$

$$\Rightarrow \quad \frac{1}{f(x)-20} \simeq a + b\,x \;.$$

Substituímos então o problema de aproximar $f(x)$ através de $\dfrac{1}{a+b\,x}+c$ pelo de aproximar $\dfrac{1}{f(x)-20}$ através de $a+b\,x$. A vantagem prática é que temos agora uma família linear nos parâmetros e podemos determiná-los resolvendo o sistema (8.15). Assim, fazendo:

$$F(x) \;=\; \frac{1}{f(x)-20} \;=\; \begin{pmatrix} 0.0154 \\ 0.0182 \\ 0.0212 \\ 0.0239 \\ 0.0266 \\ 0.0299 \end{pmatrix} \;;$$

$$g_0(x) \;=\; \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \;;\; g_1(x) \;=\; \begin{pmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 8 \\ 10 \end{pmatrix} \;;$$

e usando o produto escalar usual, obtemos:

$$\begin{pmatrix} 6 & 30 \\ 30 & 220 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0.1352 \\ 0.7748 \end{pmatrix} ,$$

cuja solução é: $a_0 = 0.0155 = a$ e $a_1 = 0.0014 = b$. Portanto:

$$f(x) \simeq \frac{1}{0.0155 + 0.0014\ x} + 20$$

Novamente minimizamos o quadrado da diferença entre a função $F(x)$ e a reta $a_0 + a_1\ x$.

Exercícios

8.12 - *Deseja-se aproximar uma função f definida em um intervalo $[a, b]$ por uma função*

$$g(x) = x^2\ ln \left(\frac{x^3}{a + b\ x^2} \right) ,$$

usando o método dos minímos quadrados.

a) *Qual é a função a ser minimizada?*

b) *Qual é o sistema linear a ser resolvido?*

8.13 - *Sejam $f(x)$ e $g(x)$, funções reais distintas não identicamente nulas. Suponha que, usando o método dos minímos quadrados, aproximamos $f(x)$ em $[a, b]$ por:*

$$F(x) = a_0\ f(x) + a_1\ g(x) .$$

a) *Quais os valores que devem ser atibuídos para a_0 e a_1?*

b) *Qual será o erro?*

8.14 - *É possível aproximar diretamente uma função $f(x)$ tabelada, por uma função do tipo:*

$$g(x) = \left(\frac{a}{1 + b\ \cos\ x} \right) ,$$

usando o método dos minímos quadrados? Se não for possível, qual é a transformação que deve ser feita?

8.15 - *Considere a função dada por:*

x	1.5	2.0	2.5	3.0
$f(x)$	2.1	3.2	4.4	5.8

a) *Ajuste os pontos acima por uma função do tipo $\sqrt{a + bx}$, usando o método dos mínimos quadrados.*

b) *Qual função foi minimizada?*

8.16 - *Ajustar os valores da tabela:*

x	2	5	8	11	14	17	27	31	41	44
$f(x)$	94.8	89.7	81.3	74.9	68.7	64.0	49.3	44.0	39.1	31.6

através de uma das famílias de funções:

$$a\ e^{bx} \ , \qquad \frac{1}{a+bx} \ , \qquad \frac{x}{a+bx} \ .$$

Use o teste de alinhamento para decidir qual das funções melhor resolve o problema.

O método dos mínimos quadrados se aplica também à determinação da melhor solução de sistemas lineares incompatíveis. Assim, passamos a descrever

8.5 Sistemas Lineares Incompatíveis

Ocorre frequentemente na prática problema da natureza seguinte: temos que determinar uma função y que depende linearmente de certas variáveis $x_1, x_2, \ldots, x_m$, isto é,

$$y \ = \ c_1 x_1 \ + \ c_2 x_2 \ + \ldots + \ c_m x_m \ ,$$

onde os $c_i, \ i = 1, 2, \ldots, m$, são coeficientes desconhecidos fixados.

Na maioria dos casos os c_i são determinados experimentalmente, perfazendo-se um certo número de medidas das grandezas $x_1, x_2, \ldots, x_m$ e y. Se designarmos por $x_{j1}, x_{j2}, \ldots, x_{jm}, y_j$ os resultados correspondentes à j-ésima medida, tentamos determinar $c_1, c_2, \ldots, c_m$ a partir do sistema de equações:

$$\left\{ \begin{array}{l} x_{11}c_1 \ + \ x_{12}c_2 \ + \ \ldots \ + \ x_{1m}c_m \ = \ y_1 \\ x_{21}c_1 \ + \ x_{22}c_2 \ + \ \ldots \ + \ x_{2m}c_m \ = \ y_2 \\ \ldots \ \ldots \\ x_{n1}c_1 \ + \ x_{n2}c_2 \ + \ \ldots \ + \ x_{nm}c_m \ = \ y_n \end{array} \right. \tag{8.16}$$

Em geral, o número n de medidas é maior que o número m de incógnitas e devido aos erros experimentais o sistema (8.16) resulta ser incompatível e sua solução só pode ser obtida aproximadamente. O problema que precisa, então, ser resolvido é o da determinação dos $c_1, c_2, \ldots, c_m$ de modo que o lado esquerdo das equações (8.16) forneça resultados tão "**próximos**" quanto possível dos correspondentes resultados do lado direito. Resta-nos apenas, para a solução do problema, precisarmos o conceito de proximidade. Como medida dessa proximidade adotaremos o quadrado da distância euclidiana entre y e z do $\mathbb{R}^n$, onde:

$$y \ = \ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \ , \ z \ = \ \begin{pmatrix} x_{11}c_1 \ + \ x_{12}c_2 \ + \ \ldots \ + \ x_{1m}c_m \\ x_{21}c_1 \ + \ x_{22}c_2 \ + \ \ldots \ + \ x_{2m}c_m \\ \ldots \ \ldots \\ x_{n1}c_1 \ + \ x_{n2}c_2 \ + \ \ldots \ + \ x_{nm}c_m \end{pmatrix} \ .$$

Assim:

$$Q \ = \| \ z - y \ \|^2 = \ \sum_{i=1}^{n} \left(x_{i1}c_1 \ + \ x_{i2}c_2 \ + \ldots + \ x_{im}c_m \ - \ y_i \right)^2 \ .$$

Esse problema, como sabemos, pode ser facilmente resolvido, determinando-se a projeção de $y \in I\!R^n$ sobre o subespaço gerado pelos vetores:

$$g_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} ; \; g_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} , \; \ldots \; , g_m = \begin{pmatrix} x_{1m} \\ x_{2m} \\ \vdots \\ x_{nm} \end{pmatrix} ,$$

uma vez que o vetor z que procuramos, tal que $\| z - y \|^2$ seja mínima, pode ser expresso como:

$$z = c_1 g_1 + c_2 g_2 + \ldots + c_m g_m .$$

Isto é, $z \in$ ao sub-espaço gerado por $g_1, g_2, \ldots, g_m$. Supondo os g_i, $i = 1, \ldots, m$, linearmente independentes, a solução do problema é dada por:

$$\begin{pmatrix} (g_1, g_1) & (g_2, g_1) & \ldots & (g_m, g_1) \\ (g_1, g_2) & (g_2, g_2) & \ldots & (g_m, g_2) \\ \ldots\ldots & & & \\ (g_1, g_m) & (g_2, g_m) & \ldots & (g_m, g_m) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix} = \begin{pmatrix} (y, g_1) \\ (y, g_2) \\ \vdots \\ (y, g_m) \end{pmatrix} , \tag{8.17}$$

onde:

$$(g_i, g_j) = \sum_{k=1}^{n} x_{ki} \, x_{kj} \; ; \; (y, g_j) = \sum_{k=1}^{n} y_k \, x_{kj} .$$

Exemplo 8.10 - *Determinar, pelo método dos mínimos quadrados, o valor de c na equação:*

$$y = cx ,$$

sabendo-se que c satisfaz às equações:

$$\begin{aligned} 2c &= 3 \\ 3c &= 4 \\ 4c &= 5 \end{aligned}$$

Solução: Temos que:

$$g_1 = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{e} \quad y = \begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix} .$$

Assim, o sistema (8.17) reduz-se a:

$$(g_1, g_1) \, c = (y, g_1)$$

onde:

$$\begin{aligned} (g_1, g_1) &= 4 + 9 + 16 = 29 , \\ (y, g_1) &= 6 + 12 + 20 = 38 , \end{aligned}$$

Logo:

$$29\ c \ = \ 38 \quad \Rightarrow \ c \ = \ \frac{38}{29} \ .$$

Portanto a reta procurada é:

$$y \ = \ \frac{38}{29}\ x.$$

Note que o caso acima é um caso particular do problema de se determinar c tal que:

$$y \ = \ cx \ .$$

O sistema (8.16) reduz-se a um sistema de n equações a uma incógnita:

$$\begin{array}{l} x_1\ c \ = \ y_1 \\ x_2\ c \ = \ y_2 \\ \vdots \\ x_n\ c \ = \ y_n \end{array}$$

Como é fácil de se verificar, c deve satisfazer:

$$c \ = \ \frac{(g_1, y)}{(g_1, g_1)} \ = \ \frac{\sum_{k=1}^{n} x_k\ y_k}{\sum_{k=1}^{n} x_k^2} \ .$$

Nesse caso, c é geometricamente interpretado como o coeficiente angular da reta através da origem, $y = cx$, que está *tão próximo quanto possível* dos pontos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Exercícios:

8.17 - *Determinar a melhor solução para o sistema:*

$$\left\{ \begin{array}{rcrcr} x_1 & - & x_2 & = & -1 \\ 2x_1 & + & x_2 & = & -2 \\ -x_1 & - & 3x_2 & = & 1 \\ 2x_1 & + & 3x_2 & = & -2 \\ 3x_1 & - & 2x_2 & = & -3 \end{array} \right.$$

8.18 - *Resolver, pelo método dos mínimos quadrados, o sistema:*

$$\left\{ \begin{array}{rcrcrcr} 3x_1 & + & x_2 & - & x_3 & = & 2 \\ x_1 & + & 2x_2 & + & x_3 & = & 3 \\ 2x_1 & - & x_2 & + & 3x_3 & = & -1 \\ -x_1 & + & x_2 & - & x_3 & = & 0 \\ -2x_1 & - & 3x_2 & - & 2x_3 & = & 2 \end{array} \right.$$

8.6 Exercícios Complementares

8.19 - *De uma tabela são extraídos os valores:*

x	-2	-1	0	1	2
y	6	3	-1	2	4

Usando o método dos mínimos quadrados ajuste os dados acima por polinômio de grau adequado.
Sugestão: use gráfico.

8.20 - *Considere a tabela:*

x	-2	-1	1	2
y	1	-3	1	9

a) *Pelo método dos mínimos quadrados, ajuste à tabela as funções:*

$$g_1(x) = ax^2 + bx; \qquad g_2(x) = cx^2 + d \ .$$

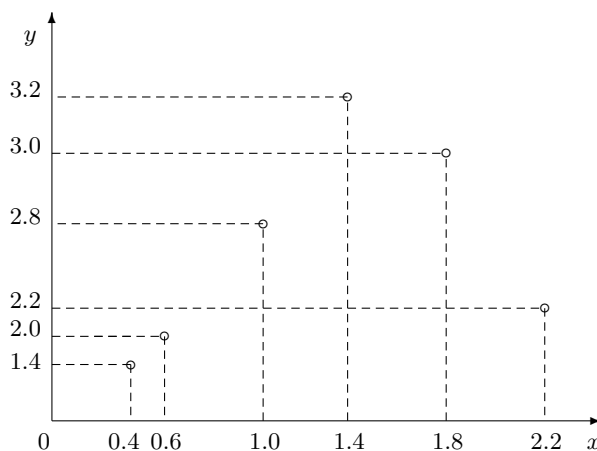
b) *Qual das funções fornece o melhor ajuste segundo o critério dos mínimos quadrados? Justifique.*

8.21 - *Achar aproximação mínimos quadrados da forma:*

$$g(x) = ae^x + be^{-x} \ ,$$

correspondente aos dados:

x_i	0	0.5	1.0	1.5	2.0	2.5
y_i	5.02	5.21	6.49	9.54	16.02	24.53

8.22 - *Um dispositivo tem uma certa característica y que é função de uma variável x. Através de várias experiências foi obtido o gráfico:*

Deseja-se conhecer o valor de y para $x = 0.5$. Da teoria sabe-se que a função que descreve y tem a forma aproximada de uma curva do tipo $a\ x^2 + b\ x$. Obtenha valores aproximados para a e b, usando todas as observações, e então estime o valor para y quando $x = 0.5$.

8.23 - *Usando método dos mínimos quadrados aproxime a função* $f(x) = 3\ x^6 - x^4$ *no intervalo [-1,1], por uma parábola, usando os polinômios de Legendre.*

Obs: Os polinômios de Legendre

$$L_n(x) = \frac{1}{2^n.n!} \quad \frac{d^n}{dx^n}(x^2-1)^n; \quad L_0(x) = 1 \ ,$$

são ortogonais segundo o produto escalar:

$$(L_i(x), L_j(x)) = \int_{-1}^{1} L_i(x) L_j(x) dx \ ,$$

e além disso satisfazem:

$$\int_{-1}^{1} {L_n}^2(x) dx = \frac{2}{2n+1} \ .$$

8.24 - *Encontre a melhor aproximação para* $f(x) = sen\ 3x$ *no intervalo* $[0, 2\pi]$, *da forma:*

$$P(x) \ = \ a_0 + a_1 \ \cos \ x + b_1 \ sen \ x \ .$$

8.25 - *Determinar aproximação trigonométrica de ordem 2, em* $[-\pi, \pi]$, *para* $y(t) = |sent|$.

8.26 - *Obter aproximação trigonométrica, de ordem 1, para a função :*

x	$\frac{\pi}{3}$	$\frac{2\pi}{3}$	π	$\frac{4\pi}{3}$	$\frac{5\pi}{3}$	2π
$f(x)$	0.75	−0.75	−6	−0.75	0.75	6

8.27 - *Deseja-se aproximar uma função* $f > 0$ *tabelada por uma função do tipo:*

$$\frac{x^2}{ln\ (a\ x^4 + b\ x^2 + c)} \ ,$$

usando o método dos minímos quadrados.

a) *Podemos aproximar f diretamente pela função acima? Caso não seja possível, quais são as transformações necessárias?*

b) *Qual função será minimizada?*

c) *Qual é o sistema linear a ser resolvido?*

8.28 - *Usando o método dos mínimos quadrados encontre a e b tais que* $y = ax^b$ *ajusta os dados:*

x	0.1	0.5	1.0	2.0	3.0
y	0.005	0.5	4	30	110

8.29 - *Considere a tabela:*

t	3.8	7.0	9.5	11.3	17.5	31.5	45.5	64.0	95.0
y	10.0	12.5	13.5	14.0	15.0	16.0	16.5	17.0	17.5

Por qual das funções :

$$a) \quad y(t) = \frac{t}{a+bt}; \qquad b) \quad y = ab^t$$

você ajustaria esta tabela?

Sugestão: Faça o teste de alinhamento.

8.30 - *Qual das funções:*

$$I) \quad y = ax^2 + b; \qquad II) \quad y = \frac{1}{a + bx} \ ,$$

ajusta melhor os valores da tabela:

x	0	0.5	1.0	1.5	2.0	2.5	3.0
y	-2	-1.5	-0.5	1.5	4.5	9.0	17.0

Usando o método dos mínimos quadrados ajuste os valores da tabela pela função escolhida.

8.31 - *Considere a tabela:*

x	1.0	1.5	2.0	2.5	3.0
y	1.1	2.1	3.2	4.4	5.8

Ajuste os pontos acima por uma função do tipo $x ln(ax + b)$, usando o método dos mínimos quadrados .

8.32 - *Físicos querem aproximar os seguintes dados:*

x	0.1	0.5	1.0	2.0
$f(x)$	0.13	0.57	1.46	5.05

usando a função $ae^{bx} + c$. Eles acreditam que $b \simeq 1$.

i) *Calcule os valores de a e c pelo método dos mínimos quadrados, assumindo que b = 1.*

ii) *Use os valores de a e c obtidos em i) para estimar o valor de b.*

Observe que neste exercício, devemos, depois de encontrado o valor de b, (item ii)), recalcular o valor de a e c e a seguir b, e assim por diante.

8.33 - *Usando o método dos mínimos quadrados achar a solução aproximada do sistema incompatível:*

$$\begin{cases} a - 3b = 0.9 \\ 2a - 5b = 1.9 \\ -a + 2b = -0.9 \\ 3a - b = 3.0 \\ a + 2b = 1.1 \end{cases}$$

8.34 - *Determine pelo método dos mínimos quadrados a melhor solução para sistema linear incompatível:*

$$\begin{cases} 2N + 3M = 8 \\ N + M = 6 \\ 3N + M = 5 \\ N + 3M = 12 \end{cases}$$

onde N e M são restritos a valores inteiros.

8.35 - *A tabela a seguir foi obtida da observação de determinado fenômeno que sabe-se é regido pela equação:*

$$E = ax + by \ .$$

Sabendo que:

x	2	1	1	-1	2
y	3	-1	1	2	-1
E	1	2	-1	1	3

i) *Determine pelo método dos mínimos quadrados a e b.*

ii) *Qual o erro cometido?*

8.7 Problemas Aplicados e Projetos

8.1 - *A tabela a seguir lista o número de acidentes em veículos motorizados no Brasil em alguns anos entre 1980 e 1998.*

ano	*nº de acidentes (em milhares)*	*acidentes por 10000 veículos*
1980	*8.300*	*1.688*
1985	*9.900*	*1.577*
1990	*10.400*	*1.397*
1993	*13.200*	*1.439*
1994	*13.600*	*1.418*
1996	*13.700*	*1.385*
1998	*14.600*	*1.415*

a) *Calcule a regressão linear do número de acidentes no tempo. Use-a para prever o número de acidentes no ano 2000. (Isto é chamada uma análise de série temporal, visto que é uma regressão no tempo e é usada para prognosticar o futuro).*

b) *Calcule uma regressão quadrática do número de acidentes por 10000 veículos. Use esta para prognosticar o número de acidentes por 10000 veículos no ano 2000.*

c) *Compare os resultados da parte* **a)** *e* **b)**. *Em qual delas você está mais propenso a acreditar?*

Observe que em qualquer trabalho de série temporal envolvendo datas contemporâneas é uma boa idéia eliminar os dados iniciais antes de formar as somas; isto reduzirá os problemas de arredondamento. Assim, em vez de usar para x os valores 1980, 1985, etc, usamos 0, 5, etc.

8.2 - *A tabela a seguir lista o total de água (A) consumida nos Estados Unidos em bilhões de galões por dia:*

ano	1950	1960	1970	1980	1990
(A)	136.43	202.70	322.90	411.20	494.10

a) *Encontre uma regressão exponencial de consumo de água no tempo.*

b) *Use os resultados do item a) para prever o consumo de água nos anos de 1995 e 2000.*

8.3 - *O número de números primos menores que x é denotado por $\Pi(x)$ e vale a tabela:*

x	100	1000	10000	100000
$\Pi(x)$	25	168	1229	9592

a) *Determinar pelo método dos mínimos quadrados , para os dados acima, expressão do tipo:*

$$\Pi(x) = a + b\frac{x}{log_{10}x} \ .$$

b) *Estimar o número de números primos de 6 dígitos usando o item* **a)**.

8.4 - *Mr. K. P. Lear (1609, Way of Astronomy) teve a idéia de que a Terra se move ao redor do Sol em órbita elíptica, com o Sol em um dos focos. Depois de muitas observações e cálculos, ele obteve a tabela a seguir, onde r é a distância da Terra ao Sol, (em milhões de Km) e x é o ângulo (em graus) entre a linha Terra-Sol e o eixo principal da elipse.*

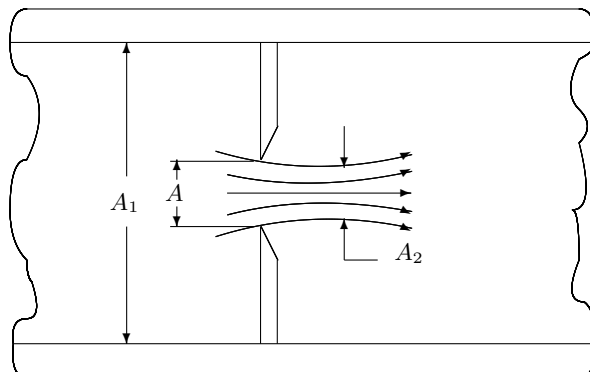
x	0	45	90	135	180
r	147	148	150	151	152

Mr. Lear sabe que uma elipse pode ser escrita pela fórmula:

$$r = \frac{\rho}{1 + \epsilon \cos\ x} \ .$$

Com os valores da tabela ele pode agora estimar os valores de ρ e ϵ. Ajude Mr. Lear a estimar os valores de ρ e ϵ, (depois de um re-arranjo da fórmula acima).

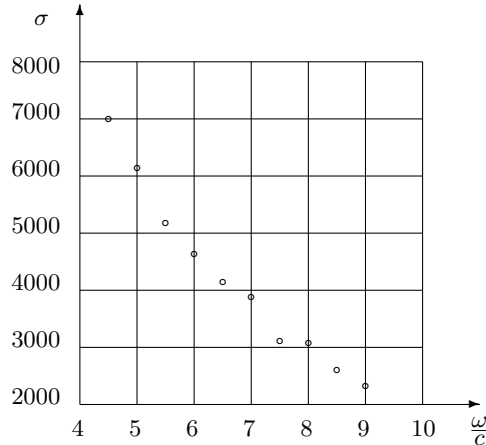
8.5 -*Placas de orifício com bordas em canto (ou faca) são muito utilizadas na medição da vazão de fluidos através de tubulações . A figura a seguir mostra uma placa de orifício, que tem os seguintes parâmetros geométricos representativos:*

- A = *área da seção reta do orifício*
- A_1 = *área da seção reta da tubulação*
- $A_2 = CA$ *(seção reta no ponto de maior contração após o orifício)*

O coeficiente C é função da razão A/A_1, e valores experimentais desse coeficiente estão listados na tabela a seguir:

A/A_1	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
C	0.62	0.63	0.64	0.66	0.68	0.71	0.76	0.81	0.89	1.00

a) *Fazendo $x = A/A_1$, ajuste a função $C(x)$ pela função: $a_0 + a_1x + a_2x^2$ aos pontos da tabela, usando o método dos mínimos quadrados.*

b) *Faça um gráfico dos valores fornecidos pelo polinômio. Acrescente a esse gráfico os valores dos pontos da tabela. Comparando visualmente a curva dos valores fornecidos pelo polinômio, e os valores da tabela, você pode concluir que a aproximação obtida é boa?*

8.6 - *A resistência à compressão do concreto, σ, decresce com o aumento da razão água/cimento, $\frac{\omega}{c}$, (em galões de água por saco de cimento). A resistência à compressão de três amostras de cilindros para várias razões $\frac{\omega}{c}$ estão mostradas no gráfico a seguir:*

e, cujos valores estão na tabela:

$\frac{\omega}{c}$	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0
σ	7000	6125	5237	4665	4123	3810	3107	3070	2580	2287

a) *Usando o método dos mínimos quadrados, ajuste σ, aos dados, utilizando uma função do tipo: $k_1 e^{-k_2 \frac{\omega}{c}}$.*

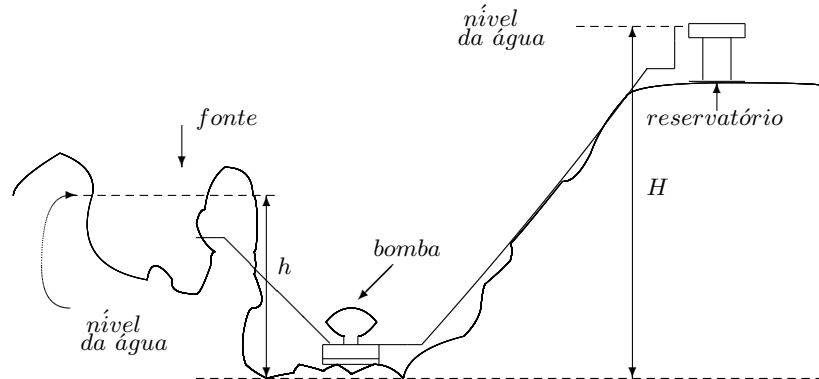
b) *Compare os valores da curva obtida no item* **a)** *com os do gráfico, para verificar (por inspeção), se a curva obtida para σ é uma boa aproximação.*

8.7 - *Um fazendeiro, verificando a necessidade de construir um novo estábulo, escolheu um local próximo a uma nascente, de forma que, perto do estábulo, pudesse também ter um reservatório de água. Junto à nascente ele construiu uma barragem e instalou uma bomba, para que a água pudesse chegar ao reservatório.*

Verificou-se que:

i) *A vazão da fonte de alimentação , Q, era aproximadamente de 30 litros por minuto. (Quantidade de água que aflui à bomba).*

ii) *A altura da queda, h, era de 6 metros. (Altura entre a bomba e o nível da água da fonte de alimentação).*

iii) *O reservatório se encontrava a uma altura de recalque, H, de 46 metros. (Altura entre a bomba e o nível da água no reservatório).*

Munidos destes dados, o fazendeiro gostaria de saber quantas vacas leiteiras poderiam ocupar o estábulo, sabendo que o consumo diário de cada uma, incluindo o asseio do estábulo, é de 120 litros.

Observação:

Para resolver o problema deve-se calcular a vazão de recalque, q, que é a quantidade de litros por minuto que entram no reservatório. Para isso tem-se que aplicar a fórmula:

$$q = Q \quad \frac{h}{H} \quad R \ ,$$

onde R é o rendimento da bomba.

Concliu-se portanto, que para determinar a vazão de recalque é necessário conhecer o rendimento da bomba. A tabela a seguir relaciona a razão entre as alturas $\frac{h}{H}$ e o rendimento da bomba instalada.

$\frac{h}{H}$	6.0	6.5	7.0	7.5	8.0	8.5	9.0
R	0.6728	0.6476	0.6214	0.5940	0.5653	0.5350	0.5020

Consultando a tabela verificou-se que para calcular o R associado a um valor de $\frac{h}{H}$ deveria ser feita uma regressão linear.

8.8 *Após serem efetuadas medições num gerador de corrente contínua, foram obtidos os seguintes valores indicados por um voltímetro e um amperímetro.*

$I(carga(A))$	1.58	2.15	4.8	4.9	3.12	3.01
$V(v)$	210	180	150	120	60	30

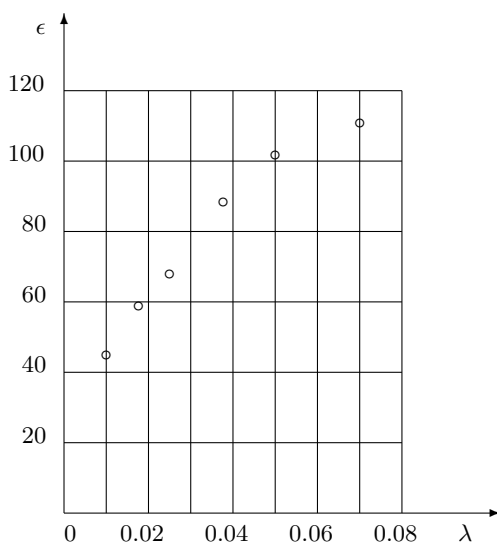
Faça um gráfico dos dados.

a) *Ajuste os dados por polinômio de grau adequado.*

a) *Estime o valor a ser obtido no voltímetro quando o amperímetro estiver marcando* $3.05A$.

8.9 - *Um tubo fino e comprido está imerso na água. Na base do tubo há um reator elétrico, que descreve um movimento sobe-e-desce oscilatório. Em qualquer instante, a coordenada y da base do tubo é dada pela fórmula:* $y = \lambda sen\ (\omega t)$*, onde* λ *é a amplitude do movimento e* ω *é a frequência de oscilação. Medidores de deformação colocados na base do tubo medem a deformação do mesmo, em função da amplitude de oscilação* λ*. Os valores da deformação* ϵ *medida em relação as amplitudes* λ *encontram-se na tabela:*

λ	0.010	0.020	0.0025	0.038	0.050	0.070
ϵ	45×10^{-6}	59×10^{-6}	69×10^{-6}	87×10^{-6}	101×10^{-6}	112×10^{-6}

Colocando os dados da tabela, num gráfico:

parece apropriado aproximar a função por uma parábola. Assim, usando o método dos mínimos quadrados:

a) *ajuste a função* ϵ *por função do tipo:* $a_0 + a_1\lambda + a_2\lambda^2$,

b) *Faça um gráfico dos valores de* $\epsilon(\lambda)$ *contra* λ*, obtidos através da parábola, e marque no mesmo gráfico dos pontos da tabela. Qualitativamente falando, o polinômio é uma boa aproximação?*

8.10 - *(Ajuste na curva tração/deformação de um tipo de aço). Feito um ensaio de tração em uma barra de um tipo de aço, em uma máquina universal de Amsler, foram obtidos os valores constantes da tabela a seguir. Deseja-se obter representações aproximadas para* $d = f(t)$.

$t(ton/m^2)$	0.8	1.8	2.8	3.8	4.8	5.8	6.8	7.8	8.8	9.8
d	0.15	0.52	0.76	1.12	147	1.71	2.08	2.56	3.19	4.35
	10.0	10.2	10.4	10.6	10.8	11.0	11.2	11.4	11.6	11.8
	4.55	5.64	6.76	8.17	10.1	12.7	16.2	20.3	30.0	60.0

Faça um gráfico dos dados. Observando o gráfico você verá que deve fazer:

a) *uma regressão linear para os 10 primeiros pontos (*$0.8 \leq t \leq 9.8$*).*

b) *uma regressão quadrática para os 7 pontos seguintes (*$10.0 \leq t \leq 11.2$*).*

c) *uma regressão linear para os 3 últimos pontos (*$11.4 \leq t \leq 11.8$*).*

8.11 - *A tabela a seguir lista o Produto Nacional Bruto (PNB) em dólares constantes e correntes. Os dólares constantes representam o PNB baseado no valor do dólar em 1987. Os dólares correntes são simplesmente o valor sem nenhum ajuste de inflação.*

ano	PNB (dólar corrente) (em milhões)	PNB (dólar constante) (em milhões)
1980	248.8	355.3
1985	398.0	438.0
1989	503.7	487.7
1992	684.9	617.8
1994	749.9	658.1
1995	793.5	674.6
1997	865.7	707.6

Estudos mostram que a melhor forma de trabalhar os dados é aproximá-los por uma função do tipo: $a\,x^b$. *Assim:*

a) *Utilize a função* $a\,x^b$ *para cada um dos PNB's no tempo.*

b) *Use os resultados da parte* **a)** *para prever os PNB's no ano 2000.*

c) *Que lhe dizem os resultados da parte* **b)** *sobre a taxa de inflação no ano 2000?*

8.12 - *Em um estudo, determinou-se que a vazão de água em uma tubulação está relacionada com o diâmetro e com a inclinação desssa tubulação (em relação à horizontal). Os dados experimentais estão na tabela a seguir:*

Experimento	Diâmetro	Inclinação	Vazão (m^3/s)
1	1	0.001	1.4
2	2	0.001	8.3
3	3	0.001	24.2
4	1	0.01	4.7
5	2	0.01	28.9
6	3	0.01	84.0
7	1	0.05	11.1
8	2	0.05	69.0
9	3	0.05	200.0

O estudo também sugere que a equação que rege a vazão da água tem a seguinte forma:

$$Q \;=\; a_0\; d^{a_1}\; S^{a_2}\;,$$

onde Q *é a vazão (em* m^3/s*),* S *é a inclinação da tubulação,* D *é o diâmetro da tubulação (em m) e* a_0, a_1 *e* a_2 *são constantes a determinar.*

***1.0cm a)** *Usando a equação acima e o método dos mínimos quadrados determine* a_0, a_1 *e* a_2 .

***1.0cm b)** *Use o resultado do item* **a)** *para estimar a vazão em m^3/s para uma tubulação com um diâmetro de $2.5m$ e uma inclinação de 0.0025.*

8.13 - *Os dados das tabelas a seguir mostram a quantidade de alcatrão e nicotina (em miligramas) para várias marcas de cigarro com e sem filtro :*

	COM FILTRO									
ALCAT.	8.3	12.3	18.6	22.9	23.1	24.0	27.3	30.0	35.9	41.5
NICOT.	0.32	0.46	1.10	1.32	1.26	1.44	1.42	1.96	2.23	2.20

	SEM FILTRO									
ALCAT.	32.5	33.0	34.2	34.8	36.5	37.2	38.4	41.1	41.6	43.4
NICOT.	1.69	1.76	1.48	1.88	1.73	2.12	2.35	2.46	1.97	2.65

i) *Calcule as regressões lineares do tipo $ax + b$ para a relação entre nicotina (y) e alcatrão (x) em ambos os casos (com e sem filtro).*

ii) *Discuta a hipótese de a (coeficiente angular) ser o mesmo nos dois casos.*

iii) *Para uma certa quantidade de alcatrão, os cigarros com filtro contém menos nicotina que os sem filtro?*

8.14 -*Em uma floresta densa, dois formigueiros inimigos, separados por um rio, travam uma longa batalha. Um dos formigueiros, visando acabar gradativamente com a espécie inimiga, dá uma cartada fatal, sequestrando a rainha. Para não ver seu formigueiro destruído, algumas formigas guerreiras se unem com as formigas engenheiras numa arriscada operação de resgate à rainha. Mas como resgatá-la?*

As formigas guias arquitetaram um plano de resgate que consistia em arremessar algumas formigas guerreiras, algumas formigas engenheiras, algumas guias e algumas operárias sobre o formigueiro inimigo. Chegando lá, elas dividiriam-se em dois grupos. Um grupo composto pelas formigas guerreiras e formigas guias, incumbiria-se de entrar no formigueiro inimigo em busca da rainha. O outro grupo, composto pelas formigas engenheiras e pelas formigas operárias, incumbiria-se de fabricar uma catapulta que os arremessasse de volta ao seu formigueiro de origem. Para que nenhum imprevisto ocorresse quando do arremesso da rainha, as formigas engenheiras foram consultadas. Elas então criaram um modelo teórico de catapulta. A catapulta consistia em utilizar um pedaço de matinho de tamanho fixo. A rainha deveria ser colocada em uma ponta, com a catapulta inclinada em um determinado ângulo θ, e arremessada numa distância exata entre os formigueiros.No entanto, para que isso ocorresse, era necessário saber qual matinho usar. Algumas formigas então começaram a realizar testes com matinho de uma mesma espécie mas com diferentes diâmetros e verificaram que, para um alcance horizontal fixo igual à distância entre os formigueiros, o ângulo de arremesso variaria com o diâmetro do matinho, segundo a seguinte tabela de dados:

diâmetro	*ângulo*
1	*40.03*
1.079	*36.68*
1.177	*34.56*
1.290	*33.21*
1.384	*31.99*
1.511	*31.33*
1.622	*31.01*
1.766	*30.46*

Para encontrar a melhor curva descrita por esses pontos as formigas engenheiras fizeram um programa que encontrava a melhor aproximação pelo método dos mínimos quadrados. Faça você o mesmo que as formigas engenheiras!

8.15 - *Um capacitor de capacitância C Farad, com carga inicial de q Coulomb está sendo descarregado através de um circuito elétrico que possui um resistor com resistência de R Ohms. Da teoria sabe-se que em um certo instante $t \geq 0$, a corrente I no circuito é dada por*

$$I = I_0 e^{-\frac{t}{RC}}$$

onde $t = 0$ é o instante em que o circuito é ligado e $I_0 = q/RC$.

Os seguintes dados experimentais foram obtidos:

$t(s)$	1	2	3	4	5	6	7	8
$I(A)$	0.37	0.14	0.056	0.0078	0.003	0.001	0.00042	0.00022

i) *Calcule os desvios quadráticos médios:*

$$DQM_1 = [\sum_{i=1}^{m} (Y_i - A - Bt_i)^2]/m \ ,$$

$$DQM_2 = [\sum_{i=1}^{m} (I_i - ae^{bt_i})^2]/m \ ,$$

onde $Y_i = lnI_i$, $A = lna$, $B = b$ e m é o número de pontos. Qual o significado desses desvios?

ii) *Qual o tempo necessário para que a corrente seja 10% da inicial?*

8.16 - *Um modelo para o crescimento da população segundo Verhulst é que a população, P, cresce no tempo de acordo com a equação diferencial:*

$$\frac{dP}{dt} = (A - BP) \quad P \ .$$

O parâmetro A é a taxa geométrica de crescimento para populações relativamentes pequenas. Entretanto, com o crescimento da população, há um efeito de retardamento ou freio causado pelo consumo do suprimento de alimentos, poluição do meio ambiente, e assim por diante. Este efeito de freio é representado pelo parâmetro B.

A solução desta equação diferencial é dada por:

$$P = \frac{A}{B \quad + \quad c \quad e^{-Nt}} \ ,$$

onde C é a constante de integração e pode ser determinada a partir da população inicial em $t = 0$. A determinação dos parâmetros A e B requer uma regressão do tipo que não pode ser linearizada. Entretanto, suponha que aproximemos:

$$\frac{dP}{dt} = \frac{P_{k+1} - P_k}{\Delta t} \ ,$$

onde P_k é a população no final do k-ésimo período de tempo; isto é, $P_k = P(k\Delta t)$ e Δt é o acréscimo do tempo.

Então a equação diferencial, torna-se:

$$P_{k+1} = (1 + A\Delta t - B\Delta t P_k)P_k \ .$$

i) *Determinar os parâmetros A e B correspondentes aos dados a seguir, provenientes do censo em um país, onde P é dado em milhões:*

ano	1930	1940	1950	1960	1970	1980	1990
P	0.63	0.76	0.92	1.1	1.2	1.3	1.5

ii) *Usando os valores de A e B encontrados no item i) faça a previsão da população no país no ano 2000.*

8.17 - *A tabela a seguir relaciona a quantidade ideal de calorias, em função da idade e do peso, para homens que possuem atividade física moderada e vivem a uma temperatura ambiente de 20° C.*

i / *p*	*25*	*45*	*65*
50	*2500*	*2350*	*1950*
60	*2850*	*2700*	*2250*
70	*3200*	*3000*	*2550*
80	*3550*	*3350*	*2800*

i) *Usando o método dos mínimos quadrados encontre expressão da forma:*

$$cal = bp + ci \ ,$$

que aproxime a tabela.

ii) *Determine a cota de calorias para um homem:*

a) *de 30 anos e 70 quilos*

b) *de 45 anos e 62 quilos*

c) *de 50 anos e 78 quilos*

usando a expressão do item **i)**.

8.18 - *Uma refinaria pode comprar dois tipos de petróleo bruto: leve ou pesado . Os custos por barril desses tipos são respectivamente 110 e 90 (unidade adotada: reais). As seguintes quantidades de gasolina (G), querosene (Q) e combustível de avião (CA) são produzidas a partir de um barril de cada tipo de petróleo:*

	G	Q	CA
LEVE	0.4	0.2	0.35
PESADO	0.32	0.4	0.2

Note que 5% e 8% do petróleo bruto, leve e pesado, são perdidos respectivamente durante o processo de refinamento. A refinaria deve entregar 100 barris de Gasolina, 40 barris de Querosene e 25 barris de Combustível para aviões, sendo que há disponibilidade de 20 mil reais para a compra de petróleo bruto leve e pesado. O objetivo é determinar tais quantidades. Assim, denotando por x a quantidade de petróleo bruto leve e por y a quantidade de petróleo bruto pesado, em barris chegamos ao sistema:

$$\begin{cases} 110x & + & 90y & = & 20000 \\ 0.4x & + & 0.32y & = & 100 \\ 0.2x & + & 0.4y & = & 40 \\ 0.35x & + & 0.2y & = & 25 \end{cases}$$

que é um sistema incompatível.

a) *Usando o método dos mínimos quadrados determine x e y.*

b) *Pode-se obter x e y como quantidades não inteiras?*

c) *Calcule o resíduo, isto é:*

$$\sum_{i=1}^{4} (b_i \; - \; a_{i1} \; x \; - a_{i2} \; y)^2 \; ,$$

onde a_{ij} são os elementos da matriz do sistema e b_i é o vetor independente.

d) *Quais as quantidades de gasolina, querosene e combustível para aviões produzidas?*

8.19 - *O calor perdido pela superfície do corpo humano é afetado pela temperatura do ambiente e também pela presença do vento. Por exemplo, a perda de calor em -5^o C acompanhada de um vento de 10 km/h é equivalente à perda de calor em -11^o C sem vento. Dada a temperatura t e a velocidade do vento v, pode - se calcular a temperatura $\bar{t}$ que na ausência do vento, tem o efeito resfriador equivalente.*

Considerando que vale a tabela:

t / v	-10	-5	0
0	-10	-5	0
10	-16	-11	-5
20	-22	-16	-10

e supondo que vale aproximadamente:

$$\bar{t} = at + bv$$

determine:

a) *Os parâmetros* a *e* b.

b) *A temperatura* $\bar{t}$ *equivalente à temperatura de* -2^o *C e velocidade do vento de 5 Km/h.*

8.20 - *A madeira com uma grande porcentagem de nós não é tão resistente quanto a madeira sem nós. Foram estabelecidas normas para determinar a relação de resistência entre uma viga com nós e uma viga sem nós. Para vigas ou pranchas os nós são medidos na face estreita da viga. A relação de resistência percentual* R *depende da largura* L *da face e do tamanho* T *do nó. Uma tabela parcial destas relações de resistência é dada a seguir.*

T \ L	75	100	125	150
12.5	85	90	125	150
25.0	72	78	82	85
37.5	57	67	73	77
50.0	35	55	64	70
65.5	18	47	56	61

a) *Encontre expressão da forma:*

$$R = a + b\,T + c\,L\,,$$

que aproxime a tabela.

b) *Determine a proporção de resistência para nós de* 45 mm *com largura de face de* 90 mm.

Capítulo 9

Programação não Linear

Este Capítulo é do Marcos Arenales

Capítulo 10

Aproximação de Funções: Métodos de Interpolação Polinomial

10.1 Introdução

A aproximação de funções por polinômios é uma das idéias mais antigas da análise numérica, e ainda uma das mais usadas. É bastante fácil entender por que razão isso acontece. Os polinômios são facilmente computáveis, suas derivadas e integrais são novamente polinômios, suas raízes podem ser encontradas com relativa facilidade, etc.

A simplicidade dos polinômios permite que a aproximação polinomial seja obtida de vários modos, entre os quais podemos citar: Interpolação, Método dos Mínimos Quadrados, Osculação, Mini-Max, etc, portanto é vantajoso substituir uma função complicada por um polinômio que a represente. Além disso temos o Teorema de Weirstrass que afirma que: *toda função contínua pode ser arbitrariamente aproximada por um polinômio.*

Veremos aqui como aproximar uma função usando **Métodos de Interpolação Polinomial**.

Tais métodos são usados como uma **aproximação** para uma função $f(x)$, principalmente, nas seguintes situações:

a) não conhecemos a expressão analítica de $f(x)$, isto é, sabemos apenas seu valor em alguns pontos $x_0, x_1, x_2, \ldots$, (esta situação ocorre muito frequentemente na prática, quando se trabalha com dados experimentais) e necessitamos manipular $f(x)$ como, por exemplo, calcular seu valor num ponto, sua integral num determinado intervalo, etc.

b) $f(x)$ é extremamente complicada e de difícil manejo. Então, às vezes, é interessante sacrificar a *precisão* em benefício da simplificação dos cálculos.

10.2 Polinômio de Interpolação

O problema geral da interpolação por meio de polinômios consiste em, dados $n+1$ números (ou pontos) distintos (reais ou complexos) $x_0, x_1, \ldots, x_n$ e $n+1$ números (reais ou complexos) $y_0, y_1, \ldots, y_n$, números estes que, em geral, são $n+1$ valores de uma função $y = f(x)$ em $x_0, x_1, \ldots, x_n$, determinar-se um polinômio $P_n(x)$ de grau no máximo n tal que

$$P_n(x_0) = y_0 \;\; ; \;\; P_n(x_1) = y_1 \; ; \ldots ; \; P_n(x_n) = y_n .$$

Vamos mostrar que tal polinômio existe e é único, na hipótese de que os pontos $x_0, x_1, \ldots, x_n$ sejam distintos.

Teorema 10.1 *-Dados $n+1$ pontos distintos $x_0, x_1, \ldots, x_n$ (reais ou complexos) e $n+1$ valores $y_0, y_1, \ldots, y_n$ existe um e só um polinômio $P_n(x)$, de grau menor ou igual a n, tal que*

$$P_n\left(x_k\right) \;=\; y_k \quad , \quad k = 0, 1, \ldots, n \; . \tag{10.1}$$

Prova: Seja:

$$P_n(x) \;=\; a_0 + a_1\, x + \ldots + a_n\, x^n \; ,$$

um polinômio de grau no máximo n, com $n+1$ coeficientes $a_0, a_1, \ldots, a_n$ a serem determinados.

Em vista de (10.1), temos:

$$\begin{cases} a_0 & + & a_1 x_0 & + & \ldots & + & a_n x_0^n & = & y_0 \\ a_0 & + & a_1 x_1 & + & \ldots & + & a_n x_1^n & = & y_1 \\ \ldots & \ldots & & & & & & & \\ a_0 & + & a_1 x_n & + & \ldots & + & a_n x_n^n & = & y_n \end{cases} \tag{10.2}$$

o qual pode ser interpretado como um sistema linear para os coeficientes $a_0, a_1, \ldots, a_n$ e cujo determinante, conhecido como determinante de Vandermonde, é dado por:

$$V \;=\; V\;\left(x_0, x_1, \ldots, x_n\right) \;=\; \begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \ldots & & & \\ 1 & x_n & \ldots & x_n^n \end{vmatrix} \; . \tag{10.3}$$

Para se calcular V, procedemos da maneira seguinte:

Consideremos a função $V(x)$ definida por:

$$V(x) \;=\; V\;\left(x_0, x_1, \ldots, x_{n-1}, x\right) \;=\; \begin{vmatrix} 1 & x_0 & \ldots & x_0^n \\ 1 & x_1 & \ldots & x_1^n \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n-1} & \ldots & x_{n-1}^n \\ 1 & x & \ldots & x^n \end{vmatrix} \; . \tag{10.4}$$

$V(x)$ é, como facilmente se verifica, um polinômio de grau menor ou igual a n. Além disso, $V(x)$ se anula em $x_0, x_1, \ldots, x_{n-1}$. Podemos, então escrever:

$$V\left(x_0, x_1, \ldots, x_{n-1}, x\right) \;=\; A\left(x - x_0\right)\;\left(x - x_1\right)\ldots\left(x - x_{n-1}\right)\; , \tag{10.5}$$

onde A depende de $x_0, x_1, \ldots, x_{n-1}$.

Para se calcular A, desenvolvemos (10.4) segundo os elementos da última linha e observamos que o coeficiente de x^n é $V(x_0, x_1, \ldots, x_{n-1})$. Logo,

$$V\left(x_0, \ldots, x_{n-1}, x\right) \;=\; V\left(x_0, \ldots, x_{n-1}\right)\left(x - x_0\right)\ldots\left(x - x_{n-1}\right)\; . \tag{10.6}$$

Substituindo x por x_n em (10.6), obtemos a seguinte fórmula de recorrência:

$$V\left(x_0, \ldots, x_{n-1}, x_n\right) \;=\; V\left(x_0, \ldots, x_{n-1}\right)\left(x_n - x_0\right)\ldots\left(x_n - x_{n-1}\right)\; . \tag{10.7}$$

De (10.3), temos que: $V(x_0, x_1) = x_1 - x_0$. Em vista de (10.7) podemos escrever:

$$V(x_0, x_1, x_2) = (x_1 - x_0)(x_2 - x_0)(x_2 - x_1) .$$

Por aplicações sucessivas de (10.7), obtemos:

$$V(x_0, x_1, \ldots, x_n) = \prod_{i>j} (x_i - x_j) .$$

Por hipótese, os pontos $x_0, x_1, \ldots, x_n$ são distintos. Assim $V \neq 0$ e o sistema (10.2) tem uma e uma só solução $a_0, a_1, \ldots, a_n$.

Vimos, então, que dados $n + 1$ pontos distintos $x_0, x_1, \ldots, x_n$ e $n + 1$ valores $f(x_0) = y_0, f(x_1) = y_1, \ldots, f(x_n) = y_n$ de uma função $y = f(x)$, existe um e um só polinômio $P_n(x)$ de grau no máximo n tal que

$$P_n(x_k) = f(x_k) \quad , \quad k = 0, 1, \ldots, n .$$

Em vista disso, temos a seguinte definição.

Definição 10.1 - *Chama-se* **polinômio de interpolação** *de uma função $y = f(x)$ sobre um conjunto de pontos distintos $x_0, x_1, \ldots, x_n$, ao polinômio de grau no máximo n que coincide com $f(x)$ em $x_0, x_1, \ldots, x_n$. Tal polinômio será designado por $P_n(f; x)$ e, sempre que não causar confusão, simplesmente por $P_n(x)$.*

Exemplo 10.1 - *Dados os pares de pontos:* $(-1, 15)$; $(0, 8)$; $(3, -1)$, *determinar o polinômio de interpolação para a função definida por este conjunto de pares de pontos.*

Solução: Temos:

$$\begin{array}{rcl} x_0 &=& -1 \,, \\ x_1 &=& 0 \,, \\ x_2 &=& 3 \,, \end{array} \qquad \text{e} \qquad \begin{array}{rcccl} y_0 &=& 15 &=& f(x_0) \,, \\ y_1 &=& 8 &=& f(x_1) \,, \\ y_2 &=& -1 &=& f(x_2) \,. \end{array}$$

Como $n = 2$, devemos determinar $P_2(x) = a_0 + a_1\, x + a_2\, x^2$, tal que $P_2(x_k) = y_k$, $k = 0, 1, 2$, isto é:

$$\begin{cases} a_0 + a_1\, x_0 + a_2\, x_0^2 = y_0 \\ a_0 + a_1\, x_1 + a_2\, x_1^2 = y_1 \\ a_0 + a_1\, x_2 + a_2\, x_2^2 = y_2 \end{cases}$$

Substituindo x_k e y_k, $k = 0, 1, 2$, obtemos:

$$\begin{cases} a_0 - a_1 + a_2 = 15 \\ a_0 = 8 \\ a_0 + 3\, a_1 + 9\, a_2 = 9 \end{cases}$$

cuja solução é: $a_0 = 8$, $a_1 = -6$ e $a_2 = 1$. Assim:

$$P_2(x) = 8 - 6\, x + x^2 ,$$

é o polinômio de interpolação para a função dada pelos pares de pontos: $(-1, 15)$; $(0, 8)$; $(3, -1)$

Observações:

a) Observe que nos pontos tabelados, o valor do polinômio encontrado e o valor da função, devem coincidir. Se os valores forem diferentes você terá cometido erros de cálculo.

b) A determinação do polinômio de interpolação por meio de solução de sistemas é muito trabalhosa, além de poder ocorrer erros de arredondamento, fazendo com que a solução obtida seja irreal. Vamos, por isso, procurar outros métodos para determinação deste polinômio.

10.3 Fórmula de Lagrange

Sejam $x_0, x_1, \ldots, x_n$ $n+1$ pontos distintos. Consideremos para $k = 0, 1, \ldots, n$, os seguintes polinômios $\ell_k(x)$ de grau n:

$$\ell_k(x) = \frac{(x-x_0)\ldots(x-x_{k-1})\,(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)\ldots(x_k-x_{k-1})\,(x_k-x_{k+1})\ldots(x_k-x_n)}. \tag{10.8}$$

É fácil verificar que:

$$\ell_k(x_j) = \delta_{kj} = \begin{cases} 0, & \text{se} \quad k \neq j\ , \\ 1, & \text{se} \quad k = j\ . \end{cases} \tag{10.9}$$

De fato: substituindo x por x_k em (10.8) vemos que se o numerador e o denominador são exatamente iguais $\Rightarrow$ $\ell_k(x_k) = 1$. Agora se substituímos x por x_j com $j \neq k$ vemos o numerador anula-se e assim $\ell_k(x_j) = 0$.

Para valores dados: $f_0 = f(x_0),\ f_1 = f(x_1),\ \ldots,\ f_n = f(x_n)$ de uma função $y = f(x)$, o polinômio:

$$P_n(x) = \sum_{k=0}^{n} f_k\ \ell_k(x)\ , \tag{10.10}$$

é de grau no máximo n e, em vista de (10.9), satisfaz :

$$P_n(x_k) = f_k \quad , \quad k = 0, 1, 2, \ldots, n.$$

Logo $P_n(x)$, assim definido, é o polinômio de interpolação de $f(x)$ sobre os pontos $x_0, x_1, \ldots, x_n$.

A fórmula (10.10) é chamada **Fórmula de Lagrange do Polinômio de Interpolação**.

Exemplo 10.2 - *Conhecendo-se a seguinte tabela:*

x	-1	0	3
$f(x)$	15	8	-1

a) *Determine o polinômio de interpolação na forma de Lagrange.*

b) *Calcule uma aproximação para $f(1)$, usando o item* **a)**.

Solução: Temos:

$$\begin{aligned} x_0 &= -1\ , \\ x_1 &= 0\ , \\ x_2 &= 3\ , \end{aligned} \qquad \text{e} \qquad \begin{aligned} f_0 &= f(x_0) = 15\ , \\ f_1 &= f(x_1) = 8\ , \\ f_2 &= f(x_2) = -1\ . \end{aligned}$$

e, portanto, $n = 2$. Assim, o polinômio de interpolação na forma de Lagrange é dado por:

$$P_2(x) \; = \; \sum_{k=0}^{2} f_k \; \ell_k(x) \; .$$

Determinemos os polinômios $\ell_k(x), \; k = 0, 1, 2$. Temos:

$$\begin{aligned}
\ell_0(x) \; &= \; \frac{(x - x_1)\,(x - x_2)}{(x_0 - x_1)\,(x_0 - x_2)} \; = \; \frac{(x - 0)\;(x - 3)}{(-1 - 0)(-1 - 3)} \; = \; \frac{x^2 - 3\;x}{4} \; , \\
\ell_1(x) \; &= \; \frac{(x - x_0)\,(x - x_2)}{(x_1 - x_0)\,(x_1 - x_2)} \; = \; \frac{(x + 1)\;(x - 3)}{(0 + 1)(0 - 3)} \; = \; \frac{x^2 - 2\;x - 3}{-3} \; , \\
\ell_2(x) \; &= \; \frac{(x - x_0)\,(x - x_1)}{(x_2 - x_0)\,(x_2 - x_1)} \; = \; \frac{(x + 1)\;(x - 0)}{(3 + 1)(3 - 0)} \; = \; \frac{x^2 + x}{12} \; .
\end{aligned}$$

Portanto:

$$\begin{aligned}
P_2(x) \; &= \; f_0 \; \ell_0(x) \; + \; f_1 \; \ell_1(x) \; + \; f_2 \; \ell_2(x) \; = \\
&= \; 15 \; \times \; \left[\frac{x^2 - 3\;x}{4}\right] \; + \; 8 \; \times \; \left[\frac{x^2 - 2\;x - 3}{-3}\right] \; - \; 1 \; \times \; \left[\frac{x^2 + x}{12}\right] \; .
\end{aligned}$$

Agrupando os termos semelhantes, segue que:

$$P_2(x) \; = \; x^2 \; - \; 6\;x \; + \; 8 \; .$$

Uma aproximação de $f(1)$ é dada por $P_2(1)$. Assim, usando o algoritmo de Briot-Ruffini, (Capítulo 3), obtemos:

	1	-6	8
1		1	-5
	1	-5	**3**

Logo:

$$f(1) \; \simeq \; P_2(1) \; = \; 3 \; .$$

Observe que podemos obter $f(1)$ efetuando o seguinte cálculo: $P_2(1) = 1^2 - 6 \times 1 + 8 = 3$. É claro que este tipo de cálculo só deve ser utilizado para obter o resultado quando resolvemos o problema a mão. O anterior além de também poder ser utilizado a mão deve ser usado em computadores.

Vimos então que para obter o valor da função num ponto não tabelado, podemos aproximar a função por seu polinômio de interpolação e através deste ter uma aproximação do valor da função no ponto. Veremos agora um **esquema prático** para calcular o valor do polinômio de interpolação num ponto (**não tabelado**) sem determinar a expressão do polinômio.

Consideremos a fórmula de Lagrange, (10.10), e a fórmula dos $\ell_k(x)$, (10.8).

Fazendo:

$$\pi_{n+1}(x) \; = \; (x - x_0)\,(x - x_1) \ldots (x - x_n) \; ,$$

podemos escrever:

$$\ell_k(x) \; = \; \frac{\pi_{n+1}(x)}{(x - x_k)\,\pi'_{n+1}\,(x_k)} \; , \tag{10.11}$$

onde: $\pi'_{n+1}(x_k)$ é a derivada de $\pi_{n+1}(x)$ avaliada em $x = x_k$.

Primeiramente, calculamos as diferenças:

$$\begin{array}{ccccc}
\underline{x - x_0} & x_0 - x_1 & x_0 - x_2 & \ldots & x_0 - x_n \\
x_1 - x_0 & \underline{x - x_1} & x_1 - x_2 & \ldots & x_1 - x_n \\
x_2 - x_0 & x_2 - x_1 & \underline{x - x_2} & \ldots & x_2 - x_n \\
\ldots & & & & \\
x_n - x_0 & x_n - x_1 & x_n - x_2 & \ldots & \underline{x - x_n} \ .
\end{array}$$

Denotamos o produto dos elementos da primeira linha por D_0, o da segunda por D_1 e assim por diante. Observe que o produto dos elementos da 1ª linha é exatamente o denominador de $\ell_0(x)$ em (10.11), o produto dos elementos da 2ª linha, o denominador de $\ell_1(x)$, etc. O produto dos elementos da diagonal principal será, obviamente, $\Pi_{n+1}(x)$ e, então, segue que:

$$\ell_k(x) = \frac{\pi_{n+1}(x)}{D_k} \ , \quad k = 0, 1, \ldots, n \ .$$

Assim, a fórmula de Lagrange se reduz a:

$$\begin{aligned}
P_n(x) &= \pi_{n+1}(x) \sum_{k=0}^{n} \frac{f_k}{D_k} \\
&= \pi_{n+1}(x) \times S \ ,
\end{aligned}$$

onde:

$$S = \sum_{k=0}^{n} \frac{f_k}{D_k} \ .$$

Portanto, podemos obter o valor do polinômio num ponto, não tabelado, através do seguinte:

Esquema Prático

k	$(x_k - x_i)\,(k \neq i)$	D_k	f_k	$\frac{f_k}{D_k}$
0	$x - x_0 \quad x_0 - x_1 \quad x_0 - x_2 \quad \ldots \quad x_0 - x_n$	$(x - x_0)\prod_{\substack{i=0\\i\neq 0}}^{n}(x_0 - x_i)$	f_0	$\frac{f_0}{D_0}$
1	$x_1 - x_0 \quad x - x_1 \quad x_1 - x_2 \quad \ldots \quad x_1 - x_n$	$(x - x_1)\prod_{\substack{i=0\\i\neq 1}}^{n}(x_1 - x_i)$	f_1	$\frac{f_1}{D_1}$
2	$x_2 - x_0 \quad x_2 - x_1 \quad x - x_2 \quad \ldots \quad x_2 - x_n$	$(x - x_2)\prod_{\substack{i=0\\i\neq 2}}^{n}(x_2 - x_i)$	f_2	$\frac{f_2}{D_2}$
$\vdots$				
n	$x_n - x_0 \quad x_n - x_1 \quad x_n - x_2 \quad \ldots \quad x - x_n$	$(x - x_n)\prod_{\substack{i=0\\i\neq n}}^{n}(x_n - x_i)$	f_n	$\frac{f_n}{D_n}$
$\pi_{n+1}(x) \;=\; (x - x_0)\,(x - x_1)\ldots(x - x_n)$				S

Note que no esquema acima , acrescentamos mais três colunas: uma com o resultado dos produtos das linhas , a próxima com o valor de f_k e finalmente a última coluna com o valor de f_k/D_k. A soma desta última coluna fornece o valor S.

Exemplo 10.3 - *Aplicar o esquema acima ao exemplo anterior, isto é, calcular $f(1)$, sabendo que:*

x	-1	0	3
$f(x)$	15	8	-1

Solução: Montamos o esquema:

k	$(x_k - x_i)$	D_k	f_k	f_k/D_k
0	$2 \quad -1 \quad -4$	8	15	$15/8$
1	$1 \quad 1 \quad -3$	-3	8	$-8/3$
2	$4 \quad 3 \quad -2$	-24	-1	$1/24$
$\pi_3(1) \;=\; -4$				$S = -3/4$

k	$(x_k - x_i)$	D_k	f_k	f_k/D_k
0	$2 \quad -1 \quad -4$	8	15	$15/8$
1	$1 \quad 1 \quad -3$	-3	8	$-8/3$
2	$4 \quad 3 \quad -2$	-24	-1	$1/24$
$\pi_3(1) \;=\; -4$				$S = -3/4$

Assim, obtemos: $P_2(1) \;=\; \pi_3(1) \;\times\; S \;=\; (-4) \;\times\; (-3/4) \;=\; 3$, e portanto $f(1) \;\simeq\; P_2(1) \;=\; 3$.

Exercícios

10.1 - *Considere a tabela:*

x	1	3	4	5
$f(x)$	0	6	24	60

a) *Determinar o polinômio de interpolação, na forma de Lagrange, sobre todos os pontos.*

b) *Calcular* $f(3.5)$.

10.2 -*Construir o polinômio de interpolação, na forma de Lagrange, para a função* $y = sen\ \pi x$, *escolhendo os pontos:* $x_0 = 0\ ;\ x_1 = \dfrac{1}{6}\ ;\ x_2 = \dfrac{1}{2}$.

10.3 - *A integral elíptica completa é definida por:*

$$K(k) = \int_0^{\pi/2} \frac{dx}{\left(1 - k^2 sen^2 x\right)^{1/2}} .$$

Por uma tabela de valores dessa integral encontramos:

$$K(1) = 1.5708;\ K(2) = 1.5719;\ K(3) = 1.5739 .$$

Determinar $K(2.5)$, *usando polinômio de interpolação, na forma de Lagrange, sobre todos os pontos.*

10.4 - *Calcular* $e^{3.1}$ *usando a Fórmula de Lagrange sobre 3 pontos e a tabela:*

x	2.4	2.6	2.8	3.0	3.2	3.4	3.6	3.8
e^x	11.02	13.46	16.44	20.08	24.53	29.96	36.59	44.70

Observe que como queremos $e^{3.1}$ *usando 3 pontos, devemos escolher 3 pontos consecutivos na vizinhança de* 3.1. *Assim temos duas opções. Ou escolhemos:* $x_0 = 2.8,\ x_1 = 3.0$ *e* $x_2 = 3.2$ *ou então:* $x_0 = 3.0,\ x_1 = 3.2$ *e* $x_2 = 3.4$. *Em ambos os casos o erro na aproximação será da mesma ordem de grandeza.*

10.5 - *Sabendo-se que* $e = 2.72, \sqrt{e} = 1.65$ *e que a equação* $x - e^{-x} = 0$ *tem uma raiz em* $[0, 1]$, *determinar o valor dessa raiz, usando a Fórmula de Lagrange sobre 3 pontos.*

10.6 - *Dar uma outra prova de unicidade do polinômio de interpolação* $P_n(f; x)$ *de uma função* $y = f(x)$ *sobre o conjunto de pontos* $x_0, x_1, \ldots, x_n$.

Sugestão: supor a existência de outro polinômio $Q_n(f; x)$ *que seja de interpolação para* f *sobre* $x_0, x_1, \ldots, x_n$ *e considerar o polinômio:*

$$D_n(x) = P_n(f; x) - Q_n(f; x).$$

10.4 Erro na Interpolação

Como vimos, o polinômio de interpolação $P_n(x)$ para uma função $y = f(x)$ sobre um conjunto de pontos distintos $x_0, x_1, \ldots, x_n$ tem a propriedade:

$$P_n(x_k) = f_k \quad , \quad k = 0, 1, \ldots, n.$$

Nos pontos $\bar{x} \neq x_k$ nem sempre é verdade que $P_n(\bar{x}) = f(\bar{x})$. Entretanto para avaliar $f(x)$ nos pontos $\bar{x} \neq x_k,\ k = 1, 2, \ldots, n$, consideramos $P_n(x)$ como uma aproximação para a função $y = f(x)$ num certo intervalo que contenha os pontos $x_0, x_1, \ldots, x_n$ e calculamos $f(\bar{x})$ através de $P_n(\bar{x})$. Perguntas que surgem são, por exemplo, as seguintes: é o polinômio de interpolação uma boa aproximação para $f(x)$? Podemos ter idéia do erro que cometemos quando substituimos $f(x)$ por $P_n(x)$? Estas e outras perguntas são respondidas quando estudamos a teoria do termo do resto. Para isso, introduziremos dois lemas, cujas demonstrações podem ser encontradas em livros de Cálculo ou Análise Matemática.

Lema 10.1 -Teorema de Rolle - *Seja $f(x)$ contínua em $[a,b]$ e diferenciável em cada ponto de (a,b). Se $f(a) = f(b)$, então existe um ponto $x = \xi, a < \xi < b$, tal que $f'(\xi) = 0$.*

Prova: Pode ser encontrada em [, 19].

Lema 10.2 -Teorema de Rolle generalizado - *Seja $n \geq 2$. Suponhamos que $f(x)$ seja contínua em $[a,b]$ e que $f^{(n-1)}(x)$ exista em cada ponto de (a,b). Suponhamos que $f(x_1) = f(x_2) = \ldots = 0$ para $a \leq x_1 < x_2 < \ldots < x_n \leq b$. Então existe um ponto $\xi, x_1 < \xi < x_n$, tal que $f^{(n-1)}(\xi) = 0$.*

Prova: Pode ser encontrada em [, 19].

Vejamos agora um teorema que nos dá expressão do termo do erro.

Teorema 10.2 - *Seja $f(x)$ contínua em $[a,b]$ e suponhamos que $f^{(n+1)}(x)$ exista em cada ponto (a,b). Se $a \leq x_0 < x_1 < \ldots < x_n \leq b$, então*

$$R_n(f;x) = f(x) - P_n(f;x) \;=\; \frac{(x-x_0)\ldots(x-x_n)}{(n+1)!}\; f^{(n+1)}(\xi), \tag{10.12}$$

onde $min\{x, x_0, x_1, \ldots, x_n\} < \xi < max\{x, x_0, x_1, \ldots, x_n\}$. O ponto ξ depende de x.

Prova: Sendo $P_n(f;x_k) = f_k$, a função $R_n(f;x) = f(x) - P_n(f;x)$ se anula em $x = x_k,\ k = 0, 1, \ldots, n$.
Seja x fixado e tal que $x \neq x_k,\ k = 0, 1, \ldots, n$.
Consideremos as funções $K(x)$ e $F(t)$, definidas por:

$$K(x) \;=\; \frac{f(x) - P_n(f;x)}{(x-x_0)\,(x-x_1)\ldots(x-x_n)}\ ,\quad x \neq x_k\ ,\ k = 0, 1, \ldots, n, \tag{10.13}$$

e

$$F(t) = f(t) - P_n(f;t) - (t-x_0)\,(t-x_1)\ldots(t-x_n)\ \ K(x). \tag{10.14}$$

A função $F(t)$ se anula nos $n+1$ pontos $t = x_0,\ t = x_1,\ \ldots,\ t = x_n$. Anula-se também em $t = x$, em virtude de (10.13). Pelo Lema 10.2, a função $F^{(n+1)}(t)$ se anula em um ponto $\xi = \xi(x)$ tal que:

$$min\,\{x, x_0, x_1, \ldots, x_n\}\ <\ \xi\ <\ max\,\{x, x_0, x_1, \ldots, x_n\}\,.$$

Calculando então $F^{(n+1)}(t)$, tendo em vista (10.14), obtemos:

$$F^{(n+1)}(t) \;=\; f^{(n+1)}(t) \;-\; (n+1)!\ K(x).$$

Então, substituindo t por ξ, segue que:

$$0 \;=\; f^{(n+1)}(\xi) \;-\; (n+1)!\ K(x).$$

Portanto:

$$K(x) \;=\; \frac{f^{(n+1)}(\xi)}{(n+1)!}. \tag{10.15}$$

Comparando (10.15) com (10.13), temos, finalmente:

$$R_n(f;x) = f(x) - P_n(f;x) = \frac{(x-x_0)\,(x-x_1)\ldots(x-x_n)}{(n+1)!} f^{(n+1)}(\xi),$$

onde $min\,\{x, x_0, x_1, \ldots, x_n\}\ <\ \xi\ <\ max\,\{x, x_0, x_1, \ldots, x_n\}$, o que demonstra o teorema.

Em vista de (10.12) podemos escrever:

$$f(x) \ = \ P_n(f;x) \ + \ R_n(f;x). \tag{10.16}$$

O termo $R_n(f;x)$ na expressão (10.16) é chamado **termo do erro** ou **erro de truncamento**. É o erro que se comete no ponto x, quando se substitui a função por seu polinômio de interpolação calculado em x.

A importância do Teorema 10.2, é mais teórica do que prática, visto que não conseguimos determinar o ponto ξ de tal modo que seja válida a igualdade em (10.12). Na prática, para estimar o erro cometido ao aproximar o valor da função num ponto por seu polinômio de interpolação, utilizamos o seguinte corolário.

Corolário 10.1 - *Seja*

$$R_n(f;x) = f(x) - P_n(f;x).$$

Se $f(x)$ e suas derivadas até ordem $n+1$ são contínuas em $[a,b]$, então

$$|R_n(f;x)| \ \leq \ \frac{|x-x_0| \ |x-x_1| \ldots |x-x_n|}{(n+1)!} \ \max_{a \leq t \leq b} \ |f^{(n+1)}(t)|. \tag{10.17}$$

Prova: A demonstração fica como exercício.

Exemplo 10.4 *Dada a tabela:*

x	0	0.1	0.2	0.3	0.4	0.5
e^{3x}	1	1.3499	1.8221	2.4596	3.3201	4.4817

calcular um limitante superior para o erro de truncamento quando avaliamos $f(0.25)$, onde $f(x) = x\ e^{3x}$ usando polinômio de interpolação do 2^o grau.

Solução: Temos, de (10.17):

$$|R_2(f;x)| \ \leq \ \frac{|x-x_0| \ |x-x_1| \ |x-x_2|}{3!} \quad \max_{x_0 \leq t \leq x_2} \ |f'''(t)| \ .$$

Como $f(t) = t\ e^{3t}$, segue que:

$$\begin{aligned} f'(t) \ &= \ e^{3t} + 3\ t\ e^{3t} \ = \ e^{3t}\ (1 + 3\ t)\ , \\ f''(t) \ &= \ 3\ e^{3t}\ (1+3\ t) + 3\ e^{3t} \ = \ 6\ e^{3t} + 9\ t\ e^{3t}\ , \\ f'''(t) \ &= \ 18\ e^{3t} + 9\ e^{3t} + 27\ t\ e^{3t} \ = \ 27\ e^{3t}\ (1+t)\ . \end{aligned}$$

Como queremos estimar o valor da função $x\ e^{3x}$ no ponto 0.25 usando polinômio do 2^o, devemos tomar 3 pontos consecutivos nas vizinhanças de 0.25. Tomando então: $x_0 = 0.2$, $x_1 = 0.3$ e $x_3 = 0.4$, obtemos que:

$$\max_{x_0 \leq t \leq x_2} |f'''(t)| \ = \ 27\ e^{3(0.4)}\ (1 + 0.4) \ = \ 125.4998$$

Estamos portanto em condições de calcular um limitante superior para o erro de truncamento. Assim:

$$\begin{aligned} |R_2(f;x)| \ &\leq \ \frac{|(0.25-0.2)| \ |(0.25-0.3)| \ |(0.25-0.4)|}{6} \ (125.4998) \\ &\simeq \ 0.0078 \ \simeq \ 8 \times 10^{-3}. \end{aligned}$$

Pelo resultado obtido, vemos que se tomarmos um polinômio do 2º para avaliar $f(0.25)$, obteremos o resultado com duas casas decimais corretas.
Observações:

a) O número de zeros depois do ponto decimal, no resultado do erro, fornece o número de casas decimais corretas que teremos na aproximação.

b) Observe que poderíamos ter tomado: $x_0 = 0.1$, $x_1 = 0.2$ e $x_3 = 0.3$. Se tomarmos esses pontos, obtemos que $|R_2(f;x) \simeq 0.0054 \simeq 5\times 10^{-3}$ o que implica que obteremos duas casas decimais corretas na aproximação. Assim tanto faz tomarmos um ponto a esquerda, e dois a direita de 0.25, ou dois pontos a esquerda e um a direita, que o erro será da mesma ordem de grandeza.

Exercícios

10.7 - *Seja* $f(x) = 7x^5 - 3x^2 - 1$.

a) *Calcular* $f(x)$ *nos pontos* $x = 0$; $x = \pm 1$; $x = \pm 2$; $x = \pm 3$ *(usar o algorítmo de Briot-Ruffini). Construir a seguir a tabela segundo os valores crescentes de* x.

b) *Construir o polinômio de interpolação para esta função sobre os pontos* -2; -1; 0; 1.

c) *Determinar, pela fórmula (10.17), um limitante superior para o erro de truncamento em* $x = -0.5$ *e* $x = 0.5$.

10.8 - *Conhecendo-se a tabela:*

x	0.8	0.9	1.0	1.1	1.3	1.5
$\cos\ x$	0.6967	0.6216	0.5403	0.4536	0.2675	0.0707

calcular um limitante superior para o erro de truncamento quando calculamos cos 1.05 *usando polinômio de interpolação sobre 4 pontos.*

10.9 - *Um polinômio* $P_n(x)$, *de grau* n, *coincide com* $f(x) = e^x$ *nos pontos* $\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}$. *Qual o menor valor de* n *que se deve tomar a fim de que se tenha:*

$$|e^x - P_n(x)| \ \leq \ 10^{-6} \quad , \quad para \quad 0 \leq x \leq 1?$$

10.5 Interpolação Linear

No caso em que se substitui a função $f(x)$ entre dois pontos a e b por um polinômio de interpolação $P_1(x)$ do 1º grau, tal que $P_1(a) = f(a)$ e $P_1(b) = f(b)$ diz-se que se fez uma **interpolação linear** entre a e b.

Neste caso, em que $n = 1$, a fórmula (10.10) se reduz, sucessivamente, a:

$$\begin{aligned}
P_1(x) &= \sum_{k=0}^{1} f_k \ell_k(x) \ = \ f_0 \cdot \ell_0(x) \ + \ f_1 \ell_1(x) \ = \\
&= f_0 \frac{x - x_1}{x_0 - x_1} \ + \ f_1 \frac{x - x_0}{x_1 - x_0} \ = \ f(a)\frac{x-b}{a-b} \ + \ f(b)\frac{x-a}{b-a} \\
&= -\frac{x-b}{b-a}\ f(a) \ + \ \frac{x-a}{b-a}\ f(b) \ .
\end{aligned}$$

Assim, vemos que $P_1(x)$ pode ser escrito na forma de determinante, isto é:

$$P_1(x) \; = \; \frac{1}{b-a} \begin{vmatrix} f(b) & x-b \\ f(a) & x-a \end{vmatrix} \; = \; \frac{1}{b-a} \begin{vmatrix} f(a) & a-x \\ f(b) & b-x \end{vmatrix} \; . \tag{10.18}$$

Se $f(x)$ é contínua em $[a,b]$ e $f''(x)$ existe em cada ponto de (a,b), temos, para $a \leq x \leq b$, que:

$$R_1(f;x) \; = \; f(x) \; - P_1(f;x) \; = \; \frac{(x-a)\;(x-b)}{2!} \; f''(\xi), \; a < \xi < b \; . \tag{10.19}$$

Podemos determinar, no caso de interpolação linear, além do resultado obtido em (10.17), o seguinte: consideremos (10.19) e suponhamos, além disso que $f''(x)$ seja contínua em $[a,b]$. O polinômio $|(x-a)(x-b)|$ atinge seu máximo, para $a \leq x \leq b$, em $x = \frac{1}{2}(a+b)$ e este máximo é $\frac{1}{4}(b-a)^2$. Podemos, então, escrever:

$$|R_1(f;x)| \; \leq \; \frac{1}{2!} \; \frac{(b-a)^2}{4} \; \max_{a \leq t \leq b} \; |f''(t)| \; , \tag{10.20}$$

ou

$$|R_1(x)| \; \leq \; \frac{1}{8} \; (b-a)^2 \; M_1 \; , \tag{10.21}$$

onde M_1 é um limitante superior para $f''(t)$ em $[a,b]$.

Exemplo 10.5 - *Usando a tabela do exemplo 10.4, calcular $f(0.25)$ através de interpolação linear e dar um limitanter superior para o erro de truncamento.*

Solução: Da tabela do exemplo anterior, temos:

$$a \; = \; 0.2 \; , \qquad \qquad f(a) \; = \; 0.3644 \; ,$$

e

$$b \; = \; 0.3 \; , \qquad \qquad f(b) \; = \; 0.7379 \; ,$$

desde que $f(x) = x \; e^{3x}$. Assim, usando (10.18), segue que:

$$\begin{aligned} P_1(0.25) \quad &= \quad \frac{1}{b-a} \begin{vmatrix} f(a) & a-x \\ f(b) & b-x \end{vmatrix} \; = \; \frac{1}{0.3-0.2} \begin{vmatrix} 0.3644 & 0.2-0.25 \\ 0.7379 & 0.3-0.25 \end{vmatrix} \\ &= \quad \frac{1}{0.1} \times 0.05 \times (0.7379 + 0.3644) \; . \end{aligned}$$

Logo:

$$P_1(0.25) \; = \; 0.5512 \; \simeq \; f(0.25) \; .$$

Agora, pelo Exemplo 10.4, temos que: $f''(t) = e^{3t} \; (6+9) \; t$. Portanto

$$\max_{a \leq t \leq b} \; |f''(t)| \; = \; 2.4596(6 + 2.7) \; = \; 21.3985 \; = \; M_1 \; .$$

Segue, de (10.21) que:

$$|R_1(x)| \; \leq \; \frac{1}{8} \; (0.3-0.2)^2 \; \times \; 21.3985 \; \simeq \; 0.02673 \; \simeq 2 \; \times \; 10^{-2} \; .$$

Isto significa que no resultado obtido para $f(0.25)$ através do polinômio de interpolação linear temos apenas uma casa decimal correta. De fato, o resultado obtido foi $f(0.25) = 0.5512$ e se calcularmos o valor da f no ponto 0.25, numa máquina de calcular, obtemos que $f(0.25) = 0.52925$.

Exercícios:

10.10 - *Sabendo-se que* $\sqrt{1.03} = 1.0149$ *e* $\sqrt{1.04} = 1.0198$,

a) *calcular* $\sqrt{1.035}$, *por interpolação linear,*

b) *dar um limitante superior para o erro de truncamento.*

10.11 - *O valor de* $\log_{10} 12.7$ *foi computado por interpolação linear sobre os pontos 12 e 13. Mostrar que o erro de truncamento é* ≤ 0.004.

10.12 - *Seja a tabela:*

x	0	0.1	0.2	0.3	0.4	0.5
e^x	1	1.11	1.22	1.35	1.49	1.65

Usando interpolação linear sobre pontos adequados:

a) *Calcular* $f0.35$ *onde* $f(x) = x^2 e^x$.

b) *Dar um limitante superior para o erro de truncamento.*

10.6 Fórmula para Pontos Igualmente Espaçados

Quando os pontos x_i, são igualmente espaçados de $h \neq 0$, isto é, $x_{i+1} - x_i = h, \; i = 0, 1, \ldots, n-1$, onde h é um número fixado, há interesse, para futuras aplicações, em se determinar uma forma do polinômio de interpolação e do erro, em termos de uma variável u, definida da seguinte maneira:

$$u = \frac{x - x_0}{h} \, . \tag{10.22}$$

Em função da variável u, temos os seguintes teoremas.

Teorema 10.3 - *Para* r *inteiro, não negativo,*

$$x - x_r = (u - r)h \, .$$

Prova: (provaremos por indução em r). Assim:

a) Para $r = 0$, temos, de (10.22), que:

$$x - x_0 = uh = (u - 0)h.$$

b) Supondo válido para $r = p$, isto é,

$$x - x_p = (u - p)h \, .$$

c) Provemos que vale também para $r = p + 1$.

Temos:

$$\begin{aligned} x - x_{p+1} &= x - x_p + x_p - x_{p+1} \\ &= x - x_p - (x_{p+1} - x_p) = (u - p)h - h \\ &= (u - p - 1)h = (u - (p + 1))h. \end{aligned}$$

Portanto o teorema vale para todo inteiro $r \geq 0$.

Teorema 10.4 *Para r e s inteiros, não negativos:*

$$x_r \ - \ x_s \ = \ (r-s)h \ .$$

Prova: A prova, por ser semelhante à do teorema anterior, fica como exercício.

Consideremos o polinômio de interpolação de $f(x)$ sobre $x_0, x_1, \ldots, x_n$, dado por (10.10), isto é:

$$P_n(x) \ = \ \sum_{k=0}^{n} f_k \ \frac{(x-x_0)\,(x-x_1)\ldots(x-x_{k-1})\,(x-x_{k+1})\ldots(x-x_n)}{(x_k-x_0)\,(x_k-x_1)\ldots(x_k-x_{k-1})\,(x_k-x_{k+1})\ldots(x_k-x_n)} \ .$$

Fazendo a mudança de variável dada por (10.22) e usando os resultados dos teoremas 10.3 e 10.4, obtemos:

$$P_n\,(x_0+uh) \ = \ \sum_{k=0}^{n} f_k \ \frac{u(u-1)\ldots(u-(k-1))(u-(k+1))\ldots(u-n)}{k(k-1)\ldots(k-(k-1))(k-(k+1))\ldots(k-n)} \ , \qquad (10.23)$$

que é a forma de Lagrange do polinômio de interpolação para argumentos x_i igualmente espaçados de $h \neq 0$.

Esta forma do polinômio de interpolação é particularmente útil na determinação de fórmulas para integração numérica de funções.

De modo análogo, substituindo $x - x_r$ por $(u-r)h$ em (10.12), obtemos:

$$R_n(x) = R_n\,(x_0+uh) \ = \ u(u-1)\ldots(u-n) \ \frac{h^{n+1}}{(n+1)!} \ f^{(n+1)} \ (\xi), \qquad (10.24)$$

onde

$$min\,(x, x_0, \ldots, x_n) \ \leq \ \xi \ \leq max\,(x, x_0, x_1, \ldots, x_n) \ .$$

Temos que:

$$f^{(n+1)}(\xi) \ = \ \frac{d^{n+1}}{dx^{n+1}} \quad f(x) \ \Bigg| \ x = \xi \quad ,$$

mas se preferirmos exprimir $f(x)$ em termos de u, teremos:

$$\frac{d^{n+1}f}{dx^{n+1}} \ = \ \frac{1}{h^{n+1}} \ \frac{d^{n+1}}{du^{n+1}} \ ,$$

e assim:

$$R_n(u) \ = \ \frac{u(u-1)\ldots(u-n)}{(n+1)!} \ \frac{d^{n+1}f}{du^{n+1}} \ \Bigg| \ u = \eta$$

onde $\eta \ = \ \dfrac{\xi - x_0}{h}$ pertence ao intervalo $(0, n)$, se supusermos os pontos $x_0, x_1, \ldots, x_n$ em ordem crescente e $x \ \in \ (x_0, x_n)$.

Como vimos, o polinômio de interpolação para $f(x)$ sobre $n+1$ pontos $x_0, x_1, \ldots, x_n$ se escreve, em termos de $u \ = \ \dfrac{x - x_0}{h}$, como:

$$P_n\,(x_0+uh) \ = \ \sum_{k=0}^{n} \lambda_k(u) \ f_k \ , \qquad (10.25)$$

onde:

$$\lambda_k(u) \ = \ \frac{u(u-1)\ldots(u-(k-1))(u-(k+1))\ldots(u-n)}{k(k-1)\ldots(k-(k-1))(k-(k+1))\ldots(k-n)} \ . \qquad (10.26)$$

Exemplo 10.6 - *Dada a tabela, do Exemplo 10.4,*

a) *calcular $f(x) = x\ e^{3x}$ no ponto $x = 0.25$ usando polinômio de interpolação sobre 3 pontos.*

b) *dar um limitante superior para o erro de truncamento.*

Solução: Inicialmente escolhemos os 3 pontos apropriados na tabela dada e a seguir construimos a tabela de $f(x) = x\ e^{3x}$. Seja então: $x_0 = 0.2,\ x_1 = 0.3$ e $x_2 = 0.4$. Assim:

x	0.2	0.3	0.4
xe^{3x}	0.3644	0.7379	1.3280

a) De (10.26), temos:

$$\lambda_0(u) = \frac{(u-1)(u-2)}{(0-1)(0-2)} = \frac{u^2 - 3\ u + 2}{2}\ ,$$

$$\lambda_1(u) = \frac{u(u-2)}{1(1-2)} = \frac{u^2 - 2\ u}{-1}\ ,$$

$$\lambda_2(u) = \frac{u(u-1)}{2(2-1)} = \frac{u^2 - u}{2}.$$

Usando (10.25), obtemos:

$$\begin{aligned} P_2\left(x_0 + uh\right) &= \sum_{k=0}^{2} f_k \lambda_k(u) \\ &= (0.3644) \times \frac{u^2 - 3\ u + 2}{2} \\ &+ (0.7379) \times \left(-u^2 + 2\ u\right) \\ &+ (1.3280) \times \frac{(u^2 - u)}{2}\ . \end{aligned}$$

Agrupando os termos semelhantes, segue que:

$$P_2\left(x_0 + uh\right) = 0.1083\ u^2 + 0.2652\ u + 0.3644\ .$$

Queremos calcular $f(0.25)$. De (10.22) temos:

$$u = \frac{x - x_0}{h} \Rightarrow u = \frac{0.25 - 0.2}{0.1} = 0.5\ .$$

Usando o algorítmo de Briot-Fuffini:

	0.1083	0.2652	0.3644
0.5		0.0542	0.1597
	0.1083	0.3194	0.5241

Então $P_2(0.5) = 0.5241 \simeq f(0.25)$.

b) De (10.24) temos:

$$R_2(u) = u(u-1)(u-2)\ \frac{h^3}{3!}\ f'''(\xi)$$

Analogamente a (10.17), podemos escrever:

$$|R_2(u)| \;\leq\; |u(|u-1)(u-2)| \frac{h^3}{3!} \max_{0.2 \leq t \leq 0.4} |f'''(t)| \; ,$$

onde: $u = 0.5, \; h = 0.1 \; \Rightarrow \; h^3 = 0.001$ e pelo exemplo 10.4, temos que:

$$f'''(t) \;=\; 27 \; e^{3t} \; (1+t) \Rightarrow \max_{0.2 \leq t \leq 0.4} |f'''(t)| \;=\; 125.4988 \; .$$

Portanto:

$$\begin{aligned} |R_2(u)| \;&\leq\; |0.5| \; |(0.5-1)| \; |(0.5-2)| \;\times\; \frac{0.001}{6} \;\times\; (5.066) \;= \\ &=\; 0.0078 \;\simeq\; 8 \times 10^{-3} \end{aligned}$$

Observações:

a) Se compararmos o valor obtido para $f(0.25)$ com o valor exato veremos que o resultado está com duas casas decimais corretas.

b) O polinômio de interpolação obtido neste exemplo está em função da variavél u. Assim não é possível verificar se o valor do polinômio nos pontos tabelados coincide com a valor da função nesses pontos. Entretanto como a função é crescente no intervalo [0.2, 0.4], o valor para $f(0.25)$ deve estar entre [0.3644, 0.7379].

Observe que quando se conhece a expressão analítica da função, o termo do resto, fornece uma estimativa sobre o número de casas decimais corretas que podemos obter na aproximação. Além disso, a aplicação da fórmula do termo do resto é útil quando queremos o resultado com uma precisão pré-fixada, como mostraremos no exemplo a seguir.

Exemplo 10.7 - *Determinar o número de pontos necessários para se obter $xe^{3x}, \; x \in [0, 0.4]$ com duas casas decimais corretas usando interpolação linear sobre pontos igualmente espaçados de h.*

Solução: Por (**??**, temos que:

$$R_1 \leq \frac{(b-a)^2}{8} M_1 \;=\; \frac{h^2}{8} M_1 \leq 0.5 \times 10^{-2} \; ,$$

desde que $b-a = h$, pois os pontos são igualmente espaçados, e o erro deve ser menor ou igual a 0.5×10^{-2}, pois queremos o resultado com duas casas decimais corretas. Agora,

$$M_1 = \max_{0 \leq t \leq 0.4} |f''(t)| = e^{3t}(6+9t) \simeq 31.873$$

Portanto

$$\begin{aligned} & \frac{h^2}{8}(31.873) \leq 0.5 \times 10^{-2} \\ \Rightarrow \quad & h^2 \leq 0.001255 \\ \Rightarrow \quad & h \leq 0.00354 \; . \end{aligned}$$

Para determinar o número de pontos basta lembrar que o intervalo dado é: [0, 0.4], e portanto o número de pontos será obtido fazendo:

$$h \;=\; \frac{0.4-0}{n} \;\Rightarrow\; n \;=\; \frac{0.4}{0.0354} \simeq 11.299 \;=\; 12 \; .$$

Observe que n assim obtido é o índice do último ponto, e como tal deve ser um inteiro. Portanto o número de pontos necessários é $n+1$, ou seja, 13.

Exercícios

10.13 - *Seja a função $f(x)$ dada pela tabela:*

x	-1	0	1
$f(x)$	-4	-1	2

Calcular $f(0.5)$ usando polinômio de interpolação para argumentos igualmente espaçados.

10.14 - *Dada a função $f(x) = 4x^5 - 2x + 2$, tabelá-la nos pontos $x = 0$; $x = \pm 1$; $x = \pm 2$ e construir o seu polinômio de interpolação no intervalo $[-2; 2]$.*

10.15 - *Determinar o único polinômio de grau menor ou igual a 3, que coincide com $f(x)$ nos seguintes pontos:*

$$f(0.5) = 2 \; ; \; f(0.6) = 8 \; ; \; f(0.7) = -2 \; ; \; f(0.8) = 5$$

Calcular também $f(0.56)$.

10.16 - *A função*

$$y = \int_x^\infty \frac{e^{-t}}{t} dt \; ,$$

é dada pela seguinte tabela:

x	0	0.01	0.02	0.03	0.04	0.05	0.06
y	∞	4.0379	3.3547	2.9591	2.6813	2.4679	2.2953

Calcular y para $x = 0.0378$ usando polinômio de interpolação sobre 4 pontos.

10.17 - *Dada a função $f(x) = xe^{x/2}$ e a tabela:*

x	2	2.25	2.5	2.75	3.0
$e^{x/2}$	2.71	3.08	3.49	3.96	4.48

a) *Calcular o polinômio de interpolação sobre todos os pontos.*

b) *Calcular $f(2.4)$.*

c) *Dar um limitante superior para o erro de truncamento.*

10.7 Outras Formas do Polinômio de Interpolação

O método de Lagrange para determinação do polinômio de interpolação de uma função $y = f(x)$ sobre um conjunto de pontos $x_0, x_1, \ldots, x_n$ possui um inconveniente. Sempre que se deseja passar de um polinômio de grau p (construído sobre $p+1$ pontos) para um polinômio de grau $p+1$ (construído sobre $p+2$ pontos) todo o trabalho tem que ser praticamente refeito. Seria interessante se houvesse possibilidade de, conhecido o polinômio de grau p, passar-se para o de grau $p+1$ apenas acrescentando-se mais um termo ao de grau p. Vamos ver, agora, que tal objetivo é alcançado através da forma de Newton do polinômio de interpolação. Para a construção do polinômio de interpolação por este método, precisamos da noção de diferença dividida de uma função.

10.7.1 Diferença Dividida

Definição 10.2 - *Sejam $x_0, x_1, \ldots, x_n$, $n+1$ pontos distintos no intervalo $[a,b]$ e sejam $f_0, f_1, \ldots, f_n$, $n+1$ valores de uma função $y = f(x)$ sobre $x = x_k$, $k = 0, 1, \ldots, n$. Define-se:*

$$f\,[x_k] \;=\; f\,(x_k) \;, \qquad k = 0, 1, \ldots, n \;;$$

$$f\,[x_0, x_1, \ldots, x_n] \;=\; \frac{f\,[x_1, x_2, \ldots, x_n] \;-\; f\,[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0} \;,$$

onde $f[x_0, x_1, \ldots, x_n]$ é a **diferença dividida** *de ordem n da função $f(x)$ sobre os pontos $x_0, x_1, \ldots, x_n$.*

Assim, usando a definição, temos que:

$$f\,[x_0, x_1] \;=\; \frac{f\,[x_1] - f\,[x_0]}{x_1 - x_0} \;,$$

$$f\,[x_0, x_1, x_2] \;=\; \frac{f\,[x_1, x_2] - f\,[x_0, x_1]}{x_2 - x_0} \;,$$

$$f\,[x_0, x_1, x_2, x_3] \;=\; \frac{f\,[x_1, x_2, x_3] - f\,[x_0, x_1, x_2]}{x_3 - x_0} \;.$$

Observe que do lado direito de cada uma das igualdades acima devemos aplicar sucessivamente a definição de diferença dividida até que os cálculos envolvam apenas o valor da função nos pontos, isto é,

$$\begin{aligned} f\,[x_0, x_1, x_2] \;&=\; \frac{f\,[x_1, x_2] - f\,[x_0, x_1]}{x_2 - x_0} \\ &=\; \frac{\dfrac{f(x_2) - f(x_1}{x_2 - x_1} - \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0} \;. \end{aligned}$$

Entretanto, podemos calcular as diferenças divididas de um função, de uma maneira mais simples.

10.7.2 Cálculo Sistemático das Diferenças Divididas.

Para calcular as diferenças divididas de uma função $f(x)$ sobre os pontos $x_0, x_1, \ldots, x_n$, construímos a tabela de diferenças divididas:

Tabela de Diferenças Divididas

x_i	$f[x_i]$	$[x_i, x_j]$	$f[x_i, x_j, x_k]$	$\dots$
x_0	$f[x_0] = f_0$			
		$f[x_0, x_1] = \dfrac{f[x_1] - f[x_0]}{x_1 - x_0}$		
x_1	$f[x_1] = f_1$		$f[x_0, x_1, x_2] = \dfrac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$	
		$f[x_1, x_2] = \dfrac{f[x_2] - f[x_1]}{x_2 - x_1}$		$\dots$
x_2	$f[x_2] = f_2$		$f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$	
		$f[x_2, x_3] = \dfrac{f[x_3] - f[x_2]}{x_3 - x_2}$		$\dots$
x_3	$f[x_3] = f_3$		$f[x_2, x_3, x_4] = \dfrac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2}$	
		$f[x_3, x_4] = \dfrac{f[x_4] - f[x_3]}{x_4 - x_3}$	$\vdots$	$\dots$
x_4	$f[x_4] = f_4$	$\vdots$		
$\vdots$	$\vdots$			

da seguinte maneira:

a) a primeira coluna é constituída dos pontos $x_k, \; k = 0, 1, \dots, n$;

b) a segunda coluna contém os valoress de $f(x)$ nos pontos $x_k, \; k = 0, 1, 2, \dots, n$;

c) nas colunas $3, 4, 5, \dots$, estão as diferenças divididas de ordem $1, 2, 3, \dots$. Cada uma dessas diferenças é uma fração cujo numerador é sempre a diferença entre duas diferenças divididas consecutivas e de ordem imediatamente inferior e cujo denominador é a diferença entre os dois extremos dos pontos envolvidos.

Exemplo 10.8 - *Para a seguinte função tabelada:*

x	-2	-1	0	1	2
$f(x)$	-2	29	30	31	62

construir a tabela de diferenças divididas.

Solução: Usando o esquema acima temos que:

x_i	$f[x_i]$	$f[x_i, x_j]$	$f[x_i, x_j, x_k]$	$f[x_i, \dots, x_\ell]$	$f[x_i, \dots, x_m]$
-2	-2				
		$\dfrac{29 - (-2)}{-1 - (-2)} = 31$			
-1	29		$\dfrac{1 - (-31)}{0 - (-2)} = -15$		
		$\dfrac{30 - 29}{0 - (-1)} = 1$		$\dfrac{0 - (-15)}{1 - (-2)} = 5$	
0	30		$\dfrac{1 - 1}{1 - (-1)} = \mathbf{0}$		$\dfrac{5 - 5}{2 - (-2)} = 0$
		$\dfrac{31 - 30}{1 - 0} = 1$		$\dfrac{15 - 0}{2 - (-1)} = 5$	
1	31		$\dfrac{31 - 1}{2 - 0} = 15$		
		$\dfrac{62 - 31}{2 - 1} = 31$			
2	62				

Assim, o elemento **0**, corresponde à diferença dividida $f[x_1, x_2, x_3]$. Portanto usando a definição, segue que: $f[x_1, x_2, x_3] = \dfrac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}$ e usando o item **c)** acima, temos que: $f[x_1, x_2, x_3] = \dfrac{1-1}{1-(-1)} = 0$.

Observação: Como veremos adiante, os resultados a serem utilizados na construção do polinômio de interpolação na forma de Newton são os primeiros valores em cada coluna de diferenças embora tenhamos que construir toda a tabela pois os valores não são independentes uns dos outros.

10.7.3 Alguns Resultados sobre Diferenças Divididas

Teorema 10.5 - *As diferenças divididas de ordem k de uma função $f(x)$, satisfazem:*

$$\begin{aligned} f[x_0, x_1, \ldots, x_k] &= \frac{f[x_0]}{(x_0 - x_1)(x_0 - x_2)\ldots(x_0 - x_k)} \\ &+ \frac{f[x_1]}{(x_1 - x_0)(x_1 - x_2)\ldots(x_1 - x_k)} \\ &+ \ldots + \frac{f[x_k]}{(x_k - x_0)(x_k - x_1)\ldots(x_k - x_k - x_{k-1})} . \end{aligned}$$

Corolário 10.2 - *As diferenças divididas de ordem k de uma função $f(x)$, satisfazem:*

$$f[x_0, x_1, \ldots, x_k] = f[x_{j_0}, x_{j_1}, \ldots, x_{j_k}] ,$$

onde $(j_0, j_1, \ldots, j_k)$ *é qualquer permutação dos inteiros* $(0, 1, \ldots, k)$.

Corolário 10.3 - *As diferenças divididas de ordem k de uma função $f(x)$, satisfazem:*

$$\begin{aligned} & f[x_0, x_1, \ldots, x_k] = \\ & = \frac{f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k] - f[x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k]}{x_j - x_i} , i \neq j . \end{aligned}$$

Observações:

a) O Corolário 10.2 afirma que a diferença dividida de $f(x)$, é uma função simétrica de seus argumentos, isto é, independe da ordem dos pontos $x_0, x_1, \ldots, x_k$.

b) O Corolário 10.3 afirma que podemos tirar quaisquer dois pontos para construir a diferença dividida de uma função, e não necessariamente o primeiro e o último.

10.7.4 Fórmula de Newton

Para obtermos a forma de Newton do polinômio de interpolação precisamos inicialmente definir algumas funções. Para tanto, consideremos que $f(x)$ seja contínua e que possua derivadas contínuas em $[a, b]$, e além disso, que os pontos $x_0, x_1, \ldots, x_n$ sejam distintos em $[a, b]$. Definimos então as funções:

(1) $f[x_0, x] = \dfrac{f[x] - f[x_0]}{x - x_0}$, definida em $[a, b]$, para $x \neq x_0$.

(2) $f[x_0, x_1, x] = \dfrac{f[x_0, x] - f[x_0, x_1]}{x - x_1}$, definida em $[a, b]$,

para $x \neq x_0$ e $x \neq x_1$.

$\vdots$

(n+1) $f[x_0, x_1, \ldots, x_n, x] = \dfrac{f[x_0, x_1, \ldots, x_{n-1}, x] \; - \; f[x_0, x_1, \ldots, x_n]}{x - x_n}$,

definida em $[a, b]$, para $x \neq x_k, \; k = 0, 1, \ldots, n$.

Observe que nas funções definidas acima acrescentamos, sucessivamente, na diferença dividida o próximo ponto da tabela. Em todas estamos aplicando o Corolário 10.3. Nosso objetivo agora é encontrar uma fórmula de recorrência para $f(x)$. Assim, de (1) temos que:

$$f(x) \; = \; f[x_0] \; + \; (x - x_0) \; f[x_0, x] \; .$$

De (2), (usando (1)), obtemos:

$$\begin{aligned} & f[x_0, x_1, x]\,(x - x_1) \; = \; f[x_0, x] \; - \; f[x_0, x_1] \\ \Rightarrow \; & f[x_0, x_1, x]\,(x - x_1) \; = \; \frac{f[x] - f[x_0]}{x - x_0} \; - \; f[x_0, x_1] \\ \Rightarrow \; & f(x) \; = \; f[x_0] + (x - x_0) \; f[x_0, x_1] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x] \; . \end{aligned}$$

De maneira análoga, de (n+1), obtemos:

$$\begin{aligned} f(x) \; = \; & \{ f[x_0] + (x - x_0)\, f[x_0, x_1] + (x - x_0)\,(x - x_1)\, f[x_0, x_1, x_2] \\ + \; & (x - x_0)\,(x - x_1)\,(x - x_2)\, f[x_0, x_1, x_2, x_3] \; + \; \ldots \\ + \; & (x - x_0)\,(x - x_1) \; \ldots \; (x - x_{n-1})\, f[x_0, x_1, \ldots, x_n] \; \}_{\mathbf{1}} \\ + \; & \{(x - x_0)\,(x - x_1) \ldots (x - x_n)\, f[x_0, x_1, \ldots, x_n, x]\}_{\mathbf{2}} \, . \end{aligned} \tag{10.27}$$

Temos assim, obtido uma fórmula de recorrência para $f(x)$. Vejamos o que significam $\{\ldots\}_{\mathbf{1}}$ e $\{\ldots\}_{\mathbf{2}}$ em (10.27).

Teorema 10.6 *O polinômio:*

$$\begin{aligned} P_n(x) \; = \; & f[x_0] + (x - x_0)\, f[x_0, x_1] \\ + \; & \ldots + (x - x_0) \ldots (x - x_{n-1})\, f[x_0, x_1, \ldots, x_n] = \{\ldots\}_1 \end{aligned}$$

é o polinômio de interpolação da função $y = f(x)$ *sobre os pontos* $x_0, x_1, \ldots, x_n$*, isto é,* $P_n(x_k) = f(x_k), \; k = 0, 1, \ldots, n$.

Prova: (provaremos por indução em n). Assim:

i) Para $n = 1$, temos que:

$$\begin{aligned} P_1(x) &= f[x_0] + (x - x_0)\ f[x_0, x_1] \\ &= f[x_0] + (x - x_0)\ \frac{f[x_1] - f[x_0]}{x_1 - x_0}\ . \end{aligned}$$

Logo:

$$\text{para } x = x_0 \quad \Rightarrow \quad P_1(x_0) = f[x_0] + (x_0 - x_0)\frac{f[x_1] - f[x_0]}{x_1 - x_0} \ = \ f[x_0]\ ,$$

$$\text{para } x = x_1 \quad \Rightarrow \quad P_1(x_1) = f[x_0] + (x_1 - x_0)\frac{f[x_1] - f[x_0]}{x_1 - x_0} \ = \ f[x_1]\ .$$

2) Suponhamos válido para $n = k - 1$, isto é, $P_{k-1}(x_i) = f(x_i),\ i = 0, 1, \ldots, k - 1$.

3) Provemos para $n = k$. Divideremos a prova em duas partes. Assim:

a) Seja $i < k$; então:

$$\begin{aligned} P_k(x_i) &= P_{k-1}(x_i) + (x_i - x_0)(x_i - x_1)\ldots(x_i - x_{k-1})\, f[x_0, x_1, \ldots, x_k] \\ &= P_{k-1}(x_i) \ = \ f(x_i)\ , \end{aligned}$$

usando a hipótese de indução.

b) Seja $i = k$; então:

$$\begin{aligned} P_k(x_k) &= f[x_0] + (x_k - x_0)\, f[x_0, x_1] + \ldots \\ &+ (x_k - x_0)(x_k - x_1)\ldots(x_k - x_{k-1})\, f[x_0, x_1, \ldots, x_k] \end{aligned}$$

Fazendo $x = x_k$ em (10.27), (lembrando que $n = k$), e comparando com a expressão obtida acima para $P_k(x_k)$, vemos que $P_k(x_k) = f(x_k)$, o que completa a prova do teorema.

Teorema 10.7 - *Para* $x \in$*[a,b]*$, x \neq x_k,\ k = 0, \ldots, n$,

$$f[x_0, x_1, x_2, \ldots, x_n, x] \ = \ \frac{f^{(n+1)}(\xi)}{(n+1)!}\ ;\ \xi \ \in \ (x_0, x_n)\,.$$

Prova: Usando o Teorema 10.6, em (10,27), podemos escrever:

$$f(x) = P_n(x) \ + \ (x - x_0)\ldots(x - x_n)\, f[x_0, x_1, \ldots, x_n, x]$$

$$\Rightarrow \quad f(x) - P_n(x) = (x - x_0)\ldots(x - x_n)\, f[x_0, x_1, \ldots, x_n, x]\ . \tag{10.28}$$

Por outro lado, usando (10.12), temos que:

$$f(x) - P_n(x) = R_n(x) = (x - x_0)(x - x_1)\ldots(x - x_n)\,\frac{f^{(n+1)}(\xi)}{(n+1)!}; \tag{10.29}$$

onde $\xi \in (x_0, x_n)$.

Mas $(x - x_0)(x - x_1) \dots (x - x_n) \neq 0$, pois os pontos tabelados são distintos, Assim comparando (10.28) com (10.29), segue que:

$$f[x_0, x_1, \dots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad ; \quad \xi \in (x_0, x_n) \ .$$

Portanto:

$$\begin{aligned} P_n(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &+ \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x_0, x_1, \dots, x_n] = \{\dots\}_{\mathbf{1}} \ , \end{aligned}$$

chama-se **Forma de Newton do Polinômio de Interpolação** ou **Fórmula de Newton**, e
$R_n(x) = (x - x_0)(x - x_1) \dots (x - x_n) f[x_0, x_1, \dots, x_n, x] = \{\dots\}_{\mathbf{2}}$,

chama-se **termo do resto** ou **erro de truncamento**.

Observe que o tratamento do erro de truncamento é, portanto, o mesmo da forma de Lagrange.

Exemplo 10.9 - *Conhecendo-se a seguinte tabela:*

x	-1	0	3
$f(x)$	15	8	-1

calcular $f(1)$*, usando polinômio de interpolação de Newton.*

Solução: Temos:

$$\begin{aligned} x_0 &= -1 \ , & f_0 &= f(x_0) = 15 \ , \\ x_1 &= 0 \ , & f_1 &= f(x_1) = 8 \ , \\ x_2 &= 3 \ , & f_2 &= f(x_2) = -1 \ , \end{aligned}$$

e portanto $n = 2$. Assim o polinômio de interpolaçào na forma de Newton é dado por::

$$P_2(x) = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \ .$$

Em primeiro lugar, construímos a tabela de diferenças divididas. Assim:

x	f(x)		
-1	**15**		
		-7	
0	8		**1**
		-3	
3	-1		

Portanto: $P_2(x) = 15 + (x+1)(-7) \ + \ (x+1)(x-0)(1)$.

Agrupando os termos semelhantes obtemos: $P_2(x) = x^2 - 6\ x + 8$.

Os valores de $f(1)$ é dado por $P_2(1)$, lembrando que este é um valor aproximado. Assim: $P_2(1) = 3 \ \simeq f(1)$.

Desde que pelo Teorema 10.1 o polinômio de interpolação é único e como o exemplo acima foi resolvido pela fórmula de Lagrange, era de se esperar o resultado obtido.

Como no caso de Lagrange, existe um esquema prático para calcular o valor do polinômio de interpolação num ponto sem determinar a expressão do polinômio. Assim:

Esquema Prático

Tomemos para exemplo o polinômio de interpolação de Newton de grau 3. Assim:

$$\begin{aligned} P_3(x) &= f\,[x_0] + (x-x_0)\,f\,[x_0,x_1] \ + \ (x-x_0)\,(x-x_1)\,f\,[x_0,x_1,x_2] \ + \\ &+ \ (x-x_0)\,(x-x_1)\,(x-x_2)\,f\,[x_0,x_1,x_2 x_3]\,. \\ &= f\,[x_0] + (x-x_0)\,\{f\,[x_0,x_1] + (x-x_1)\,\{f\,[x_0,x_1,x_2] + (x-x_2)\,f\,[x_0,x_1,x_2,x_3]\}\} \ . \end{aligned}$$

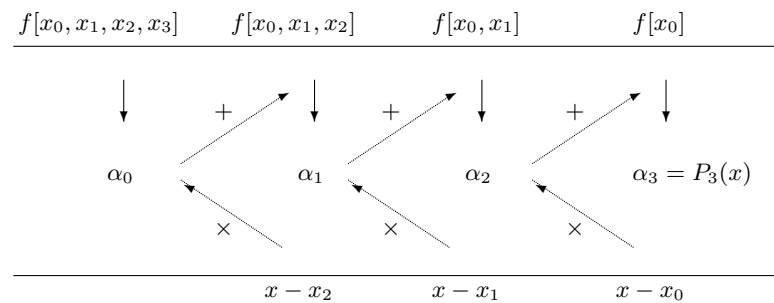
Observe que a idéia do esquema prático é ir colocando os termos comuns, que aparecem de uma determinada parcela em diante, em evidência. Denominando:

$$f\,[x_0,x_1,x_2,x_3] \ = \ \alpha_0$$

$$f\,[x_0,x_1,x_2,] \ + \ (x-x_2)\,\alpha_0 \ = \alpha_1$$

$$f\,[x_0,x_1] \ + \ (x-x_1)\,\alpha_1 \ = \alpha_2$$

$$f\,[x_0] \ + \ (x-x_0)\,\alpha_2 \ = \ \alpha_3 \ = \ P_3(x) \ .$$

O esquema prático é então dado por:

Assim para um polinômio de grau n teremos $\alpha_n = P_n(x) \simeq f(x)$.

Exemplo 10.10 - *Aplicar o esquema acima ao exemplo anterior, isto é, calcular $f(1)$.*

Solução: Temos:

$$f(x_0) = 15 , \qquad x_0 = -1 ,$$

$$f[x_0, x_1] = -7 , \qquad x_1 = 0 ,$$

$$f[x_0, x_1, x_2] = 1 , \qquad x_2 = 3 .$$

Montamos o esquema:

Assim $P_2(1) = 3 \simeq f(1)$.

Observação: Como a diferença dividida $f[x_0, x_1, \ldots, x_n, x]$ não depende da ordem de seus argumentos, podemos reescrever os pontos $x_0, x_1, \ldots, x_n, x$ em ordem crescente $x'_0, x'_1, \ldots, x'_n, x'_{n+1}$. Então pelo Teorema 10.7, temos:

$$f\left[x'_0, x'_1, \ldots, x'_n, x'_{n+1}\right] = \frac{f^{(n+1)}(\xi)}{(n+1)!} \quad ; \quad \xi \in \left(x'_0, x'_{n+1}\right) . \tag{10.30}$$

Este resultado nos permite avaliar o comportamento da derivada de ordem $n+1$ de uma função $y = f(x)$ (supondo que ela existe), por meio das diferenças divididas de ordem $n+1$ dessa função no intervalo $[a, b]$.

Em particular, a diferença dividida de ordem n de um polinômio $P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$ é independente do ponto x e igual a a_n (coeficiente de seu termo de grau n). As diferenças de ordem maior que n são todas iguais a zero. Assim, ao examinarmos uma tabela de diferenças divididas de uma função, se as diferenças de ordem k são praticamente constantes, isto significa que a função é bastante próxima de um polinômio de grau k. Podemos, então, usar um polinômio de grau k para interpolar tal função.

Exemplo 10.11 - *Dada a tabela:*

x	2	3	4	5	6	7
$f(x)$	0.13	0.19	0.27	0.38	0.51	0.67

determinar:

a) *o polinômio de interpolação de grau adequado,*

b) *calcular $f(4.5)$,*

c) *dar uma estimativa para o erro de truncamento.*

Solução: Inicialmente construímos a tabela de diferenças divididas. Assim:

x_i	$f(x_i)$	$f[x_i, x_j]$	
2	0.13		
		0.06	
3	0.19		0.01
		0.08	
4	0.27		0.015
		0.11	
5	0.38		0.01
		0.13	
6	0.51		0.015
		0.16	
7	0.67		

a) Como as diferenças divididas de 2ª ordem são praticamente constantes podemos adotar um polinômio de 2º grau para interpolá-la. Além disso, como queremos avaliar $f(4.5)$, escolhemos 3 pontos na vizinhança de 4.5. Seja então: $x_0 = 4$, $x_1 = 5$ e $x_2 = 6$. Assim:

$$\begin{aligned} P_2(x) &= f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &= 0.27 + (x - 4)(0.11) + (x - 4)(x - 5)(0.01) \\ &= 0.01\, x^2 + 0.02\, x + 0.03\,. \end{aligned}$$

b) Agora, $f(4.5) \simeq P_2(4.5) = 0.01\ (4.5)^2 + 0.02\ (4.5) + 0.03 = 0.3225$.

c) Para dar uma estimativa para o erro de truncamento calculamos as diferenças divididas de ordem 3 para a função dada. Fazendo os cálculos observamos que elas são em módulo iguais a $\frac{0.005}{3}$. Assim, usando (10.30), segue que:

$$|R_2(4.5)| \leq |(4.5 - 4)\ (4.5 - 5)\ (4.5 - 6)|\ |\frac{0.005}{3}|\ .$$

Portanto: $R_2(4.5) \simeq 0.000625 \simeq 6 \times 10^{-4}$. Logo podemos dizer que o resultado acima, obtido para $f(4.5)$, possui 3 casas decimais corretas.

Exercícios

10.18 - *Seja a função tabelada:*

x	-2	-1	1	2
$f(x)$	0	1	-1	0

a) *Determinar o polinômio de interpolação de Newton.*

b) *Calcular $f(0.5)$.*

10.19 - *Dada a função tabelada:*

x	0	1	1.5	2.5	3.0
$f(x)$	1.0	0.5	0.4	0.286	0.25

a) *Determinar o polinômio de interpolação de Newton sobre 2 pontos (interpolação linear).*

a) *Determinar o polinômio de interpolação de Newton sobre 3 pontos (interpolação quadrática).*

b) *Calcular $f(0.5)$, usando o item* **a)** *e o item* **b)**.

Lembre-se que o polinômio de Newton sobre 3 pontos é igual ao polinômio sobre 2 pontos adicionado ao termo de ordem 2. Além disso, o ponto x_0 deve ser comum aos 2 polinômios. Portanto tome cuidado ao escolher os pontos.

10.20 - *Considerando a função $f(x) = \sqrt{x}$ tabelada:*

x	1.00	1.10	1.15	1.25	1.30
$f(x)$	1.000	1.048	1.072	1.118	1.140

a) *Determinar o valor aproximado de $\sqrt{1.12}$ usando polinômio de interpolação de Newton sobre 3 pontos.*

b) *Calcular um limitante superior para o erro.*

10.21 - *Sabendo-se que a equação $x^4+6x^2-1 = 0$ tem uma raiz em [0,1], determinar o valor aproximado dessa raiz usando polinômio de interpolação de Newton sobre 3 pontos.*

10.22 - *Dada a tabela:*

x	*1*	*1.01*	*1.02*	*1.03*	*1.04*	*1.05*
$\sqrt{x}$	*1*	*1.005*	*1.01*	*1.0149*	*1.0198*	*1.0247*

a) *Calcular $\sqrt{1.035}$ por meio de um polinômio de interpolação de grau adequado.*

b) *Dar uma estimativa para o erro de truncamento.*

10.7.5 Diferenças Ordinárias

Do mesmo modo que no caso de Lagrange, existe uma fórmula mais simples para o polinômio de interpolação quando os argumentos x_i são igualmente espaçados. Consideremos então a construção do polinômio de interpolação quando os argumentos x_i são igualmente espaçados de, digamos, $h \neq 0$.

Para tanto, precisamos da noção de diferença ordinária de uma função. Assim:

Definição 10.3 - *Sejam $x_0, x_1, \ldots, x_n$, $n+1$ pontos distintos em $[a,b]$ tais que $x_{i+1} - x_i = h, i = 0, 1, \ldots, n-1$ e sejam $f_0, f_1, \ldots, f_n, n+1$ valores de uma função $y = f(x)$ sobre $x = x_k$, $k = 0, \ldots, n$. Define-se:*

$$\begin{aligned} \Delta^0 f(x_k) &= f(x_k) \ , \\ \Delta^r f(x_k) &= \Delta^{r-1} f(x_k + h) \ - \ \Delta^{r-1} f(x_k) \ , \end{aligned} \tag{10.31}$$

onde $\Delta^r f(x_k)$ é a **diferença ordinária** *de $f(x)$ de ordem r em $x = x_k$.*

Assim, usando a definição, temos:

$$\begin{aligned}
\Delta^0 f(x_k) &= f(x_k) \ , \\
\Delta^1 f(x_k) &= \Delta^0 f(x_k+h) - \Delta^0 f(x_k) = f(x_k+h) - f(x_k) \ , \\
\Delta^2 f(x_k) &= \Delta^1 f(x_k+h) - \Delta^1 f(x_k) \\
&= \Delta^0 f(x_k+2h) - \Delta^0 f(x_k+h) - \Delta^0 f(x_k+h) + \Delta^0 f(x_k) \\
&= f(x_k+2h) - 2\, f(x_k+h) + f(x_k) \ , \\
\Delta^3 f(x_k) &= f(x_k+3h) - 3f(x_k+2h) + 3f(x_k+h) - f(x_k) \ , \\
&\vdots \\
\Delta^r f(x_k) &= \begin{pmatrix} r \\ 0 \end{pmatrix} f(x_k+rh) - \begin{pmatrix} r \\ 1 \end{pmatrix} f(x_k+(r-1)h) \\
&+ \ldots + (-1)^r \begin{pmatrix} r \\ r \end{pmatrix} f(x_k) \ .
\end{aligned}$$

Portanto:

$$\begin{aligned}
\Delta^r f(x_k) &= \sum_{i=0}^{r} (-1)^i \begin{pmatrix} r \\ i \end{pmatrix} f(x_k+(r-i)h) \, ; \\
\text{onde} \quad \begin{pmatrix} r \\ p \end{pmatrix} &= \frac{r!}{p!(r-p)!} \ .
\end{aligned}$$

Entretanto, podemos calcular as diferenças ordinárias de um função, de uma maneira mais simples, através do

10.7.6 Cálculo Sistemático das Diferenças Ordinárias

Para calcular as diferenças ordinárias de uma função $f(x)$ sobre os pontos $x_0, x_1, \ldots, x_n$, (igualmente espaçados de h), construímos a

Tabela das Diferenças Ordinárias

x	$f(x)$	Δ^1	Δ^2	$\ldots$
x_0	$\Delta^0 f(x_0) = f_0$			
		$\Delta^1 f(x_0) = \Delta^0 f(x_1) - \Delta^0 f(x_0)$		
x_1	$\Delta^0 f(x_1) = f_1$		$\Delta^2 f(x_0) = \Delta^1 f(x_1) - \Delta^1 f(x_0)$	
		$\Delta^0 f(x_2) - \Delta^0 f(x_0) = \Delta^1 f(x_1)$		$\ldots$
x_2	$\Delta^0 f(x_2) = f_2$		$\Delta^2 f(x_1) = \Delta^1 f(x_2) - \Delta^1 f(x_1)$	
		$\Delta^1 f(x_2) = \Delta^0 f(x_3) - \Delta^0 f(x_2)$		$\ldots$
x_3	$\Delta^0 f(x_3) = f_3$		$\Delta^2 f(x_2) = \Delta^1 f(x_3) - \Delta^1 f(x_2)$	
		$\Delta^1 f(x_3) = \Delta^0 f(x_4) - \Delta^0 f(x_3)$	$\vdots$	
x_4	$\Delta^0 f(x_4) = f_4$	$\vdots$		
$\vdots$	$\vdots$			

da seguinte maneira:

a) a primeira coluna é constituída dos pontos $x_i,\ i = 0, 1, \ldots, n$;

b) a segunda coluna contém os valoress de $f(x)$ nos pontos $x_i,\ i = 0, 1, 2, \ldots, n$;

c) nas colunas $3, 4, 5, \ldots$, estão as diferenças ordinárias de ordem $1, 2, 3, \ldots$. Cada uma dessas diferenças é simplesmente a diferença entre duas diferenças ordinárias consecutivas e de ordem imediatamente inferior.

Exemplo 10.12 - *Para a seguinte função tabelada:*

x	-2	-1	0	1	2
$f(x)$	-2	29	30	31	62

construir a tabela de diferenças ordinárias.

Solução: Usando o esquema acima temos que:

x_k	$\Delta^0 f(x_k)$	$\Delta^1 f(x_k)$	$\Delta^2 f(x_k)$	$\Delta^3 f(x_k)$	$\Delta^4 f(x_k)$
-2	-2				
		$29-(-2)=31$			
-1	29		$1-(-31)=-30$		
		$30-29=1$		$0-(-30)=30$	
0	30		$1-1=\mathbf{0}$		$30-30=0$
		$31-30=1$		$31-1=30$	
1	31		$31-1=30$		
		$62-31=31$			
2	62				

Assim, o elemento **0**, corresponde à diferença ordinária $\Delta^2 f(x_1)$. Portanto usando a definição, segue que: $\Delta^2 f(x_1) = \Delta^1 f(x_2) - \Delta^1 f(x_1)$ e usando o item **c)** acima, temos que: $\Delta^2 f(x_1) = 1-1 = 0$.

Observação: Como veremos adiante, os resultados a serem utilizados na construção do polinômio de interpolação, para argumentos igualmente espaçados de h, são os primeiros valores em cada coluna de diferenças embora tenhamos que construir toda a tabela pois, novamente, os valores não são independentes uns dos outros.

A relação entre as diferenças divididas de ordem n de uma função e as diferenças ordinárias de $f(x)$ de ordem n num ponto x_0 é dada pelo seguinte:

Teorema 10.8 *Se* $x_k = x_0 + kh$; $k = 0, 1, \ldots, n$, *então*

$$f[x_0, x_1, \ldots, x_n] = \frac{\Delta^n f(x_0)}{h^n \cdot n!} .$$

Prova: (provaremos por indução em n). Assim:

1) para $n = 1$. Temos por definição, que:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + h) - f(x_0)}{h} = \frac{\Delta^1(x_0)}{h} ,$$

desde que $x_1 = x_0 + h$, $f(x_1) = \Delta^0 f(x_1)$ e $f(x_0) = \Delta^0 f(x_0)$.

2) Suponhamos válido para $n = k - 1$.

3) Provemos para $n = k$. Usando a definição e a seguir a hipótese de indução, obtemos:

$$\begin{aligned} f\left[x_0, x_1, \ldots, x_k\right] &= \frac{f\left[x_1, x_2, \ldots, x_k\right] - f\left[x_0, x_1, \ldots, x_{k-1}\right]}{x_k - x_0} \\ &= \frac{1}{kh}\left[\frac{\Delta^{k-1} f\left(x_1\right)}{h^{k-1}(k-1)!} - \frac{\Delta^{k-1} f\left(x_0\right)}{h^{k-1}(k-1)!}\right] = \\ &= \frac{1}{h^k\ k!}\left[\Delta^{k-1} f\left(x_0+h\right) - \Delta^{k-1} f\left(x_0\right)\right] \\ &= \frac{\Delta^k f\left(x_0\right)}{h^k \cdot k!} . \end{aligned}$$

Observe que as diferenças ordinárias de ordem n de um polinômio de grau n, $P_n(x) = a_n x^{n-1} + \ldots + a_1 x + a_0$ são iguais a $n!\ h^n \cdot a_n$. As diferenças ordinárias de ordem maior que n são todas nulas.

10.7.7 Fórmula de Newton-Gregory

Assim, usando o Teorema 10.8 no Teorema 10.6, obtemos que o polinômio de interpolação na forma de Newton, para uma função $y = f(x)$ no intervalo $[x_0, x_n]$ pode ser escrito, no caso de argumentos x_i igualmente espaçados de h, da seguinte maneira:

$$\begin{aligned} P_n(x) &= f\left(x_0\right) + \left(x-x_0\right)\frac{\Delta^1 f\left(x_0\right)}{h} + \left(x-x_0\right)\left(x-x_1\right)\frac{\Delta^2 f\left(x_0\right)}{h^2 \cdot 2!} \\ &+ \ldots + \left(x-x_0\right)\left(x-x_1\right)\ldots\left(x-x_{n-1}\right)\frac{\Delta^n f\left(x_0\right)}{h^n \cdot n!} . \end{aligned} \tag{10.32}$$

Esta forma do polinômio de interpolação é conhecida como fórmula de **Newton-Gregory**.

Exemplo 10.13 - *Dada a função tabelada:*

x	*-1*	*0*	*1*	*2*
f(x)	*3*	*1*	*-1*	*0*

determinar o polinômio de interpolação de Newton-Gregory.

Solução: Temos:

$$\begin{aligned} x_0 &= -1 , & f\left(x_0\right) &= 3 , \\ x_1 &= 0 , & f\left(x_1\right) &= 1 , \\ x_2 &= 1 , & f\left(x_2\right) &= -1 , \\ x_3 &= 2 , & f\left(x_3\right) &= 0 , \quad \text{e portanto } n = 3 . \end{aligned}$$

Assim, devemos construir o polinômio:

$$\begin{aligned} P_3(x) &= f(x_0) + (x - x_0)\,\frac{\Delta f(x_0)}{h} + (x - x_0)(x - x_1)\,\frac{\Delta^2 f(x_0)}{h^2 \cdot 2!} + \\ &+ (x - x_0)(x - x_1)(x - x_2)\,\frac{\Delta^3 f(x_0)}{h^3 \cdot 3!}\ . \end{aligned}$$

Construímos inicialmente a tabela de diferenças ordinárias. Assim:

x	f(x)			
-1	**3**			
		-2		
0	1		**0**	
		-2		**3**
1	-1		3	
		1		
2	0			

Logo:

$$\begin{aligned} P_3(x) &= 3 + (x+1)(-2) + (x+1)(x-0)\,\frac{(0)}{2} + (x+1)(x-0)(x-1)\frac{(3)}{6} \\ &= 3 - 2x - 2 + \left(x^3 - x\right)\,\frac{1}{2} \\ \Rightarrow \qquad & P_3(x) = \frac{x^3}{2} - \frac{5}{2}x + 1\ . \end{aligned}$$

Exercícios

10.23 - *Considere a função $y = f(x)$ dada pela tabela:*

x	-1	0	1	2
$f(x)$	-2	0	2	4

Determinar o polinômio de interpolação usando:

a) *forma de Newton.*

b) *forma de Newton-Gregory.*

10.24 - *Dada a função $y = sen\ x$ tabelada:*

x	1.2	1.3	1.4	1.5
$sen\ x$	0.932	0.964	0.985	0.997

a) *Calcular o polinômio de interpolação de Newton.*

b) *Calcular o polinômio de interpolação de Newton-Gregory.*

c) *Calcular sen* 1.35.

d) *Dar um limitante superior para o erro.*

10.25 - *Dada a tabela:*

0	1	2	3	4	5	6
-1	α	5	β	7	γ	13

Calcular α, β *e* γ, *sabendo-se que ela corresponde a um polinômio do 3º grau.*

Sugestão: Construa a tabela de diferenças ordinárias.

10.26 - *Dada a tabela:*

x	-2	-1	0	1
f(x)	15	0	-1	0

Calcular $f(0.5)$ *usando polinômio de interpolação sobre todos os pontos.*

10.8 Exercícios Complementares

10.27 - *Considere a função* $f(x)$ *dada pela tabela:*

x	0	1	2	3
$f(x)$	0	0	0	0

e o polinômio dado por: $P(x) = x(x-1)(x-2)(x-3)$.

a) *Verifique que:* $P(x_k) = f(x_k),\ k = 0, 1, 2, 3.$

b) $P(x)$ *é o polinômio interpolador da* $f(x)$ *sobre os pontos 0, 1, 2 e 3? Justifique.*

10.28 - *Quando conhecemos os valores de uma função y(x) e de sua derivada y'(x) em pontos dados, isto é, dados:*

$$(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \quad ,$$

$$(x_0, y'_0), (x_1, y'_1), \ldots, (x_n, y'_n) \quad ,$$

podemos montar um único polinômio $P_{2n+1}(x)$ *de grau* $\leq 2n+1$ *tal que*

$$P_{2n+1}(x_i) = y_i \quad P'_{2n+1}(x_i) = y'_i, \quad i = 0, 1, \ldots, n.$$

Sabendo que :

$$(x_0, y_0) = (0,0), \quad (x_1, y_1) = (1,1),$$

$$(x_0, y'_0) = (0,1), \quad (x_1, y'_1) = (1,0) \ ,$$

determine $P_3(x)$ *e* $P'_3(x)$ *tal que:*

$$P_3(0) = 0, \quad P_3(1) = 1, \quad P'_3(0) = 1, \quad P'_3(1) = 0 \ .$$

10.29 - *Mostre que a interpolação de um polinômio de grau n por n+k pontos, $k \geq 1$, é exata.*

10.30 - *A raiz de uma função pode ser aproximada pela raiz do seu polinômio de interpolação. Use uma parábola para determinar a raiz da função tabelada a seguir,*

x	1	2	3	4	5	6
$f(x)$	0.841	0.909	0.141	-0.757	-959	-0.279

10.31 - *Use uma parábola para determinar uma aproximação para a única raiz positiva da equação* $4\ \cos x - e^x = 0$.

10.32 - *Se $f(x)$ é um polinômio de grau 5, que característica especial tem uma tabela de diferenças divididas desse polinômio sobre 10 pontos?*

10.33 - *Frequentemente acontece que valores tabelados de uma variável y dependente de uma variável x são dados, e pretendemos achar o valor de $\bar{x}$ da variável independente correspondente a um dado $\bar{y}$ da variável dependente. Isto é conhecido como* **interpolação inversa**. *A partir da tabela:*

x	0.5	0.7	1.0	1.2	1.5	1.6
$f(x)$	-2.63	-2.57	-2.00	-1.23	0.63	0.79

determinar a raiz de f(x) usando interpolação inversa sobre 3 pontos.

10.34 - *Sabe-se que $f(x) = 5x^3 - 3x^2 + 2x - 2$ tem um zero no intervalo [0,1]. Usando interpolação inversa sobre uma tabela de 3 pontos, determinar aproximadamente $\bar{x}$ correspondente a $f(\bar{x}) = 0$.*

10.35 - *Uma maneira de se calcular o valor da derivada de uma função em um ponto x_0, quando não se conhece a expressão analítica da mesma, é usar uma tabela para formar um polinômio que aproxime a função, derivar então esse polinômio e avaliar sua derivada em $x = x_0$. Dada a tabela:*

x	0.35	0.40	0.45	0.50	0.55	0.60	0.65
$f(x)$	-1.52	1.51	1.49	1.47	1.44	1.42	1.39

calcule um valor aproximado para $f'(0.50)$ usando polinômio de interpolação de grau 2.

10.36 - *Deseja-se obter $e^x \cos x$, para $x \in [0, 1.2]$, com duas casa decimais corretas, através de interpolação linear usando uma tabela para argumentos x_i igualmente espaçados de h. Quantos valores deve ter essa tabela?*

10.37 - *A função distribuição acumulada é definida por:*

$$f(x) \ = \ \int_{-\infty}^{x} \phi(x)\ dx \text{ com } \phi(x) \ = \ \frac{1}{\sqrt{2\pi}}\ e^{-\dfrac{x^2}{2}} \ .$$

Para $0 < x < 1$ as derivadas de $f(x)$ são limitadas como segue:

$$0 < \quad f''(x) < 0.4 \ ,$$

$$0 < \quad f'''(x) < 0.5 \ ,$$

$$0.4 < \quad f^{iv}(x) < 1.2 \ .$$

Se $f(x)$ é dada com 4 casas decimais para $x = 0,\ 0.1,\ \ldots,\ 0.9,\ 1.0$, qual o grau mínimo que deve ter um polinômio interpolador se queremos 4 casas decimais precisas na aproximação de $f(x)$ para $0 < x < 1$?

10.38 - *Com que grau de precisão podemos calcular* e^{15} *usando interpolação sobre os pontos:* $x_0 = 10,\ x_1 = 20,\ x_2 = 30$ *?*

10.39 - *Com quantas casas decimais precisas podemos calcular* $\cosh 0.68$ *usando interpolação linear na tabela:*

x	0.30	0.40	0.50	0.60	0.70
$\cosh x$	1.0453	1.0811	1.1276	1.1855	1.2552

Lembre-se que : $\cosh x = \dfrac{e^x + e^{-x}}{2}$.

10.40 - *Dada a tabela:*

x	0.1	0.2	0.3	0.4	0.5
lnx	-2.303	-1.609	-1.204	-0.916	-0.693

a) *Estimar ln(0.32) através de interpolação linear e quadrática.*

b) *Qual deve ser o valor de h, se queremos obter* lnx*, com 3 casas decimais corretas, para* $x \geq 1$*, através de interpolação linear usando uma tabela para argumentos* x_i *igualmente espaçados de* h *?*

10.41 - *Considere as seguintes tabelas para uma mesma função:*

i)

x	0	1.1	2.6	3.4	4.5
$f(x)$	-1	10	13	15	24

ii)

x	0	1.1	2.6	3.4	4.5	5.8
$f(x)$	-1	10	13	15	24	34

a) *Deseja-se obter o polinômio interpolador para a tabela* **i)** *e depois para a tabela* **ii)***, de modo a fazer o menor número de operações. Qual o método ideal? Justifique.*

b) *Calcule os polinômios interpoladores para as tabelas* **i)** *e* **ii)** *usando o método escolhido no item* **a)**.

10.42 - *Suspeita -se que a tabela:*

x	-3.0	-2.0	-1.0	0.0	1.0	2.0
y	-9.0	0.0	1.0	0.0	3.0	16.0

representa um polinômio cúbico. Como testar esse fato? Explique.

10.43 - *Considere os pontos igualmente espaçados:*

$$x_0,\ x_1,\ x_2,\ x_3,\ x_4; \quad x_k = x_0 + kh,\ k = 0, 1, \ldots, 4,$$

e as diferenças divididas de 1ª *ordem de uma função* $f(x)$*, sobre esses pontos, dadas por:*

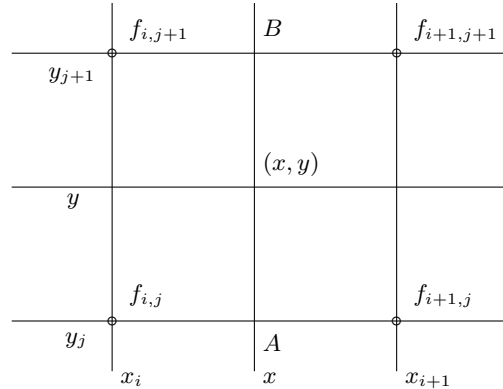
$$f[x_0, x_1] = \beta, \quad f[x_1, x_2] = \beta + 2h,$$

$$f[x_2, x_3] = \beta + 4h, \quad f[x_3, x_4] = \beta + 6h; \quad \beta \neq 0\ .$$

Qual é o grau do polinômio interpolador ? Por que?

10.44 - *Seja $f(x, y)$ uma função definida sobre os pares (x,y), com*

$$x_i \leq x \leq x_{i+1}; \quad y_j \leq y \leq y_{j+1} \ ,$$

onde $f_{r,s} = f(x_r, y_s)$.

A função $f(x, y)$ pode ser aproximada da seguinte maneira:

"Primeiro faz-se a interpolação linear através de $f_{i,j}$ e $f_{i+1,j}$ obtendo-se a aproximação f_A e em seguida, através de $f_{i,j+1}$ e $f_{i+1,j+1}$, obtendo-se a aproximação f_B. Então interpola-se linearmente através de f_A e f_B para obter a aproximação final $f(x, y)$."

a) *Seja:*

$$\alpha = \frac{x - x_i}{x_{i+1} - x_i}; \quad \beta = \frac{y - y_j}{y_{j+1} - y_j} \ .$$

Mostre que a fórmula resultante do processo acima é dada por:

$$\begin{aligned} f(x, y) \quad &= \quad (1 - \alpha)(1 - \beta) f_{i,j} + \alpha(1 - \beta) f_{i+1,j} \\ &+ \quad (1 - \alpha)\beta f_{i,j+1} + \alpha\beta f_{i+1,j+1}. \end{aligned}$$

b) *Considere a tabela para $f(x, y)$:*

y \ x	75	100	125	150
42.5	89	90	91	93
65.0	72	78	82	87
81.5	54	68	72	79
100.0	35	55	64	70
120.5	13	45	51	61

Obtenha aproximação para $f(110, 98)$.

10.9 Problemas Aplicados e Projetos

10.1 - *O gráfico de uma função f é quase um segmento de parábola atingindo seu extremo valor em um intervalo (x_0, x_2). Os valores funcionais $f_i = f(x_i)$, são conhecidos em abcissas equidistantes x_0, x_1, x_2. O valor extremo é procurado.*

a) *Use interpolação quadrática para obter a coordenada x do extremo.*

b) *Os comprimentos dos dias em Lulea, na Suécia são dados por:*

$$1 de junho: \quad 20h\ 56min$$

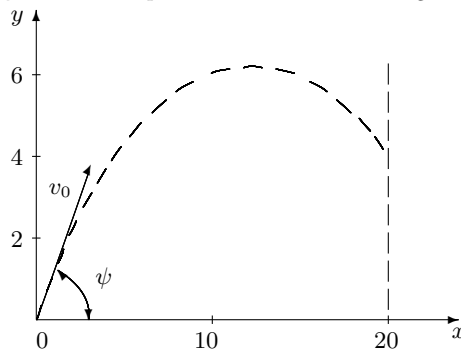
$$16 de junho: \quad 22h\ 24min$$

$$1 de julho: \quad 22h\ 01min$$

$$16 de julho: \quad 20h\ 44min$$

Use o resultado da parte **a)** *para determinar qual é o dia mais longo em Lulea e qual é sua duração.*

c) *Estime o erro cometido em* **b)**.

10.2 - *Um projétil foi lançado de um ponto tomado como origem e fez-se as seguintes observações:*

isto é:

i) *fotografou-se o projétil a 10 metros do ponto de lançamento e foi determinado sua altitude no local: 6 metros.*

ii) *uma barreira a 20 metros do ponto de lançamento interceptou-o e aí foi determinada sua altitude: 4 metros.*

Com esses 3 pontos é possível interpolar a trajetória do projétil. Comparando a equação teórica da trajetória com a obtida pela interpolação é possível determinar os parâmetros de lançamento: o ângulo ψ com a horizontal, e a velocidade inicial v_0. Assim:

a) *Determine o polinômio interpolador.*

b) *Determine ψ e v_0 sabendo que a equação da trajetória é dada por:*

$$y \ = \ x \ tg \ \psi - \frac{1}{2} \ g \ \frac{x^2}{v_0^2 \ \cos^2 \psi},$$

onde $g = 9.86 m/s^2$.

c) *Calcule a altitude do projétil a 5 metros do ponto de lançamento.*

10.3 - *Na tabela a seguir está assinalado o posicionamento de um ônibus, partindo do marco zero de uma rodovia federal,*

tempo(min)	60	80	100	120	140	160	180
posição (Km)	76	95	112	138	151	170	192

Pede - se os possíveis posicionamentos do ônibus para os tempos de 95 min., 130 min. e 170 min. Use reta e parábola.

10.4 - *Um paraquedista realizou seis saltos, saltando de alturas distintas em cada salto. Foi testada a precisão de seus saltos em relação a um alvo de raio de 5 metros de acordo com a altura. A distância apresentada na tabela a seguir é relativa à circunferência.*

Altura(m)		*Dist. do alvo(m)*
1º *salto*	*1500*	*35*
2º *salto*	*1250*	*25*
3º *salto*	*1000*	*15*
4º *salto*	*750*	*10*
5º *salto*	*500*	*7*

Levando em consideração os dados acima, a que provável distância do alvo cairia o paraquedista se ele saltasse de uma altura de 850 metros? Use reta e parábola.

10.5 - *Os resultados da densidade da água ρ em várias temperaturas estão contidos na tabela a seguir:*

T	0	5	10	15	20	25	30	35	40
ρ	0.9999	0.9998	0.9997	0.9991	0.9982	0.9971	0.9957	0.9941	0.9902

Calcular :

a) $\rho(13)$,

b) $\rho(27)$,

usando parábolas de 2º e 3º graus.

10.6 - *Conhecendo-se o diâmetro e a resistividade de um fio cilíndrico verificou-se a resistência do fio de acordo com o comprimento. Os dados obtidos estão indicados a seguir:*

Comp.(m)	500	1000	1500	2000	2500	3000	3500	4000
Res.(Ohms)	2.74	5.48	7.90	11.00	13.93	16.43	20.24	23.52

Determine quais serão as prováveis resistências deste fio para comprimentos de:

a) *1730 m,*

b) *3200 m,*

usando parábolas de 2º *e* 3º *graus.*

10.7 - *Sendo 200 candelas a intensidade de uma lâmpada, foi calculada a iluminação em casos de incidência normal, sobre uma superfície situada a distâncias conhecidas, quando para cada distância foi calculada a iluminação, conforme a tabela:*

$Dist.(metros)$	1.00	1.25	1.50	1.75	2.00	2.25	2.50
$Ilum.(lux)$	200.00	128.00	88.39	65.30	50.00	39.50	32.00

Calcular a iluminação, quando a superfície estiver situada a:

a) *1.60 m da lâmpada,*

b) *2.38 m da lâmpada,*

usando parábolas de 2º *e* 3º *graus.*

10.8 - *Um veículo de fabricação nacional, após vários testes, apresentou os resultados a seguir, quando analisou-se o consumo de combustível de acordo com a velocidade média imposta ao veículo. Os testes foram realizados em rodovia em operação normal de tráfego, numa distâcia de* $72Km$*.*

$Veloc.(Km/h)$	55	70	85	100	115	130
$Cons.(Km/l)$	14.08	13.56	13.28	12.27	11.30	10.40

Verificar o consumo aproximado para o caso de serem desenvolvidas as velocidades de:

a) *80 km/h,*

b) *105 Km/h,*

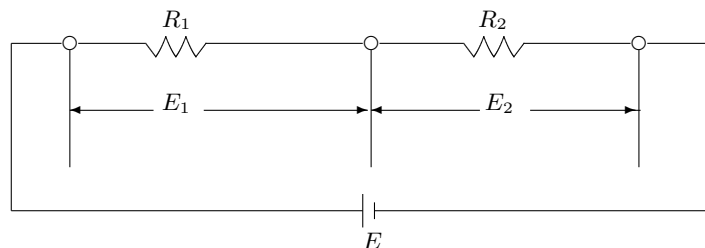
usando parábolas de 2º *e* 3º *graus.*

10.9 - *A lei de Ohm diz que:*

$$E \;=\; R\,I\;,$$

onde E *é a voltagem,* I *é a corrente e* R *a resistência, isto é, o gráfico de* $E \times I$ *é uma reta de coeficiente angular* R *que passa pela origem. Vários tipos de resistores, entretanto, não possuem essa propriedade linear. Tal resistor é chamado de um resistor não linear, ou um VARISTOR. Muitos tubos de vácuo são varistores. Geralmente, a relação entre a corrente e a voltagem para um varistor pode ser aproximada por um polinômio da forma:*

$$I \;=\; a_1 E + a_2 E^2 + \ldots + a_n E^n \;=\; \sum_{i=1}^{n} a_i E^i + P_n(E)\;.$$

Considere agora um circuito consistindo de um resistor linear R_1 *e um varistor* R_2*, como na figura a seguir:*

As equações para o circuito são:

$$I = \frac{E_1}{R_1} = \sum_{i=1}^{n} a_i E_2^i \; ;$$

$$E = E_1 + E_2 \; ,$$

e portanto,

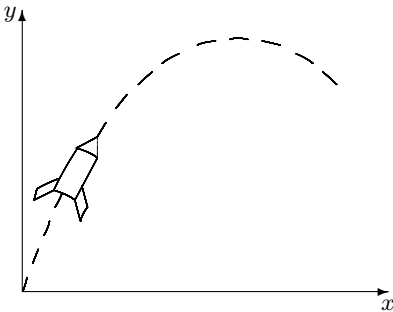
$$R_1 \left(\sum_{i=1}^{n} a_i E_2^i \right) + E_2 = E \quad \text{ou} \quad R_1 P_n(E_2) + E_2 = E \; .$$

Suponha que num certo experimento, os seguintes dados foram obtidos:

$E_2(Volts)$	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5
I(Ampéres	0.0	0.0125	0.06	0.195	0.5	1.0875	2.1	3.71	6.12	9.5625

a) *Calcule a voltagem total E e a voltagem E_1 quando $E_2 = 2.3$ Volts e $R_1 = 10$ Ohms, usando polinômio de interpolação sobre todos os pontos.*

b) *Use interpolação inversa de grau 2 para calcular a tensão no varistor quando $E_1 = 10$ Volts e $R_1 = 10 Ohms$.*

10.10 - *Um foguete é lançado na direção mostrada na figura:*

e as coordenadas x e y nos vários instantes de tempo t após o lançamento, estão dados na tabela:

t (segundos)	x (mil pés)	y (mil pés)
0	0	0
100	80	300
200	200	700
300	380	1200
400	500	1000
500	550	600

a) *Calcule* $x(250), y(250)$ *e* $y(x(250))$, *usando polinômio de interpolação sobre todos os pontos.*

b) *Compare os valores de* $y(250)$ *e* $y(x(250))$. *Os resultados são os mesmos? Deveriam ser?*

Observe que se você estiver fazendo um programa para resolver este problema, no item **a)** *você deverá interpolar* $(t_i, x_i), (t_i, y_i)$ *e* (x_i, y_i), *ou seja, existirá três polinômios interpoladores e apenas uma subrotina.*

10.11 - *A constante de equilíbrio para amônia reagindo com gases de hidrogênio e nitrogênio depende da proporção molar de hidrogênio - nitrogênio, da pressão e da temperatura. Para uma proporção molar de hidrogênio - nitrogênio 3 para 1 a constante de equilíbrio para uma faixa de pressões e temperaturas é dada por:*

Pres. / Temp.	100	200	300	400
400	0.0141	0.0159	0.0181	0.0207
450	0.0072	0.0080	0.0089	0.0103
500	0.0040	0.0044	0.0049	0.0054
550	0.0024	0.0026	0.0028	0.0031

Determinar a constante de equilíbrio para:

a) 462^o *C e uma pressão de 217 atm,*

b) 523^o *C e uma pressão de 338 atm.*

usando interpolação linear.

Observe que vale a mesma observação do problema anterior, ou seja, para cada item existirá três polinômios interpoladores e apenas uma subrotina .

10.12 - *A tabela a seguir relaciona a quantidade ideal de calorias em função da idade e do peso, para homens que possuem atividade física moderada e vivem a uma temperatura ambiente de* 20^o *C.*

idade / peso	25	45	65
50	2500	2350	1900
60	2850	2700	2250
70	3200	3000	2750
80	3550	3350	2850

Usando interpolação linear, determinar a cota aproximada de calorias para um homem

a) *de 35 anos que pesa 62 quilos,*

b *de 50 anos que pesa 78 quilos.*

Vale a mesma observação do problema anterior.

Capítulo 11

Integração Numérica

11.1 Introdução

Integrar numericamente uma função $y = f(x)$ num dado intervalo $[a, b]$ é integrar um polinômio $P_n(x)$ que aproxime $f(x)$ no dado intervalo.

Em particular, se $y = f(x)$ for dada por uma tabela ou, o que é o mesmo, por um conjunto de pares ordenados $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$ (onde os x_i podem ser supostos em ordem crescente), $x_0 = a,\ x_n = b$, podemos usar como polinômio de aproximação para a função $y = f(x)$ no intervalo $[a, b]$ o seu polinômio de interpolação.

Em particular, o polinômio de interpolação para a função $y = f(x)$ no intervalo $[a, b]$, $a = x_0,\ b = x_n$ é um polinômio de aproximação para $f(x)$ em qualquer sub-intervalo $[x_i, x_j], 0 \leq i \leq n, 0 \leq j \leq n$ do intervalo $[a, b]$. Podemos então usar o polinômio $P_n(x)$ para integrar $f(x)$ em qualquer desses sub-intervalos.

As vantagens de se integrar um polinômio que aproxima $y = f(x)$ ao invés de $f(x)$ são principalmente as seguintes:

a) $f(x)$ pode ser uma função de difícil integração ou de integração praticamente impossível, por exemplo,

$$\int_0^t \frac{s}{(t^{\frac{3}{2}} - s^{\frac{3}{2}})^{\frac{2}{3}}}\ ds\ ,$$

enquanto que um polinômio é sempre de integração imediata.

b) se conhece a solução analítica do resultado da integral, mas seu cálculo só pode ser obtido aproximadamente, por exemplo:

$$\begin{aligned} \int_0^x \frac{dt}{1+t^4} &= \frac{1}{4\sqrt{2}}\ log\ \frac{x^2 + x\sqrt{2+1}}{x^2 - x\sqrt{2+1}} \\ &+ \frac{1}{2\sqrt{2}} \left[arc\ tg \frac{x}{\sqrt{2} - x}\ +\ arc\ tg \frac{x}{\sqrt{2} + x} \right]\ . \end{aligned}$$

c) a função é dada simplesmente através de uma tabela-conjunto de pares ordenados obtidos como resultados de experiências. Aí não se conhece a expressão analítica da função em termos do argumentos x.

As fórmulas de integração são de manejo fácil e prático e nos permite, quando a função $f(x)$ é conhecida, ter uma idéia do erro cometido na integração numérica, como veremos mais adiante.

Consideraremos integrais da forma:

$$\int_a^b \omega(x) f(x) dx \ ,$$

onde $\omega(x) \geq 0$ e contínua em $[a, b]$.

A função $\omega(x)$ é chamada **função peso** e é igual a zero somente num número finito de pontos.

Usaremos **Fórmulas de Quadratura** para aproximar a integral.

Fórmulas de quadratura são aquelas que aproximam a integral usando combinação linear dos valores da função, isto é:

$$\int_a^b \omega(x) f(x) dx \ \simeq \ \sum_{k=0}^{n} A_k f(x_k) \ .$$

Seja $R(f)$ o erro, isto é:

$$R(f) \ = \ \int_a^b \omega(x) f(x) dx - \sum_{k=0}^{n} A_k f\left(x_k\right) \ .$$

Definição 11.1 - *O* **grau de precisão** *de uma fórmula de quadratura é por definição o maior inteiro* **m** *tal que* $R\left(x^k\right) = 0, \ k = 0, 1, \ldots, m$ *e* $R\left(x^{m+1}\right) \neq 0$.

Observe que isto é equivalente a dizer que a fórmula de quadratura tem grau de precisão m se é exata para todo polinômio de grau $\leq m$ e é não exata para polinômios de grau $m + 1$.

11.2 Fórmulas de quadratura interpolatória

Sejam $x_0, x_1, \ldots, x_n, \ n + 1$ pontos distintos em $[a, b]$ e sejam $f_0, f_1, \ldots, f_n, \ n + 1$ valores de uma função $y = f(x)$ sobre $x_0, x_1, \ldots, x_n$.

Seja $P_n(x)$ o polinômio de interpolação da função $y = f(x)$ sobre os $n + 1$ pontos. Pela fórmula de Lagrange, (Capítulo 10), temos que:

$$P_n(x) \ = \ \sum_{k=0}^{n} f_k \ell_k(x) \ .$$

Agora, sabemos que: $f(x) \simeq P_n(x)$ e

$$f(x) = P_n(x) + \underbrace{R_n(x)}_{\text{erro na interpolação.}} .$$

Portanto, integrando a igualdade anterior de a até b, com função peso $\omega(x)$, segue que:

$$\begin{aligned} \int_a^b \omega(x) f(x)\, dx &= \int_a^b \omega(x) P_n(x)\, dx + \underbrace{\int_a^b \omega(x) R_n(x)\, dx}_{\text{R(f)}} \\ &= \int_a^b \omega(x) \sum_{k=0}^{n} f_k \ell_k(x)\, dx + R(f) , \end{aligned}$$

onde $R(f)$ é o erro na integração.

Logo:

$$\int_a^b \omega(x) f(x)\, dx = \sum_{k=0}^{n} f_k \int_a^b \omega(x) \ell_k(x)\, dx + R(f) , \tag{11.1}$$

ou:

$$\int_a^b \omega(x) f(x)\, dx \cong \sum_{k=0}^{n} A_k f_k , \tag{11.2}$$

onde:

$$A_k = \int_a^b \omega(x) \ell_k(x)\, dx .$$

Portanto (11.2) é uma fórmula de quadratura interpolatória e

$$R(f) = \int_a^b \omega(x) R_n(x)\, dx ,$$

é o resto.

Teorema 11.1 - *A fórmula de quadratura (11.2) é interpolatória se e somente se o grau de precisão é pelo menos n. (ou seja, se e somente se a fórmula é exata para todo polinômio de grau $\leq n$).*

Prova: A prova deste teorema pode ser encontrada, por exemplo, em [Krilov, 1962].

Esse teorema garante que: dados $n+1$ pontos distintos, $x_0, x_1, \ldots, x_n$, se exigirmos que a fórmula seja exata para todo polinômio de grau $\leq n$ então os coeficientes A_k são determinados completamente. Isto é equivalente a dizer que a equação:

$$\int_a^b \omega(x)\, x^i\, dx = \sum_{k=0}^{n} A_k x_k^i ,$$

é satisfeita para $i = 0, 1, \ldots, n$ e é não satisfeita para $i = n+1$.

Exemplo 11.1 - *Seja $[a,b]$ = [0,2] e sejam $x_0 = 0, x_1 = 1$, e $x_2 = \frac{3}{2}$.*

a) *Determinar fórmula de quadratura que seja exata para todo polinômio de grau ≤ 2.*

b) *Usando a fórmula obtida em* **a)** *calcular:*

$$\int_0^2 (x^2 - 2)\ dx\ .$$

Solução: Dado os pontos x_0, x_1, x_2 exigimos que a fórmula seja exata para $f(x) = 1$; $f(x) = x$; $f(x) = x^2$. Consideremos, por simplicidade, mas sem perda de generalidade, $\omega(x) = 1$. Portanto:

$$\begin{aligned}
\int_0^2 dx &= A_0 + A_1 + A_2 \ = \ 2\ , \\
\int_0^2 x dx &= A_0 \cdot 0 \ + \ A_1 \cdot 1 \ + \ A_2 \cdot \frac{3}{2} \ = \ 2\ , \\
\int_0^2 x^2 dx &= A_0 \cdot 0 \ + \ A_1 \cdot 1 \ + \ A_2 \cdot \frac{9}{4} \ = \ \frac{8}{3}\ ,
\end{aligned}$$

onde os valores: 2, 2, $\frac{8}{3}$ são o resultado da integral de 0 a 2 de 1, x, x^2, respectivamente. Assim, temos obtido um sistema linear de 3 equações a 3 incógnitas. Resolvendo o sistema resultante, obtemos:

$$A_0 = \frac{4}{9}\ ;\ A_1 = \frac{2}{3}\ ;\ A_2 = \frac{8}{9}\ .$$

Portanto, temos uma fórmula para integrar uma função $f(x)$, no intervalo $[0, 2]$, isto é:

$$\int_0^2 f(x)dx = \frac{4}{9} f(x_0) + \frac{2}{3} f(x_1) + \frac{8}{9} f(x_2)\ ,$$

que é uma fórmula de quadratura interpolatória.

b) Temos que: $f(x) = x^2 - 2$, $x_0 = 0$, $x_1 = 1$, $x_2 = \frac{3}{2}$. Assim:

$$f(x_0) = -2\ , \quad f(x_1) = -1\ , \quad f(x_2) = \frac{1}{4}\ .$$

Portanto, usando a fórmula obtida no item **a)**, temos que:

$$\int_0^2 \left(x^2 - 2\right) dx = \frac{4}{9}(-2) + \frac{2}{3}(-1) + \frac{8}{9}\left(\frac{1}{4}\right) = -\frac{12}{9} = -\frac{4}{3}\ ,$$

e resolvendo a integral, via cálculo, obtemos:

$$\int_0^2 \left(x^2 - 2\right) dx = \left.\frac{x^3}{3} - 2\ x\right]_0^2 = \frac{8}{3} - 4 = -\frac{4}{3}\ .$$

Observe que podemos obter uma fórmula de quadratura, usando o método descrito acima. Entretanto, obter fórmulas assim é um tanto trabalhoso, pois se mudarmos os limites de integração e (ou) os pontos, todos os cálculos devem ser refeitos. Seria interessante obtermos fórmulas que não dependessem nem dos pontos nem dos limites de integração. Esse objetivo será alcançado com as

11.2.1 Fórmulas de Newton-Cotes

Estudaremos aqui **Fórmulas de Newton-Cotes do tipo fechado**. Tais fórmulas são aquelas em que a e b são pontos da fórmula de quadratura, isto é: $a = x_0$ e $b = x_n$; os argumentos x_k, são igualmente espaçados de uma quantidade fixa h, isto é, $x_{k+1} - x_k = h$, $k = 0, 1, \ldots, n-1$, a função peso, $\omega(x)$, é constante e igual a 1, e o intervalo de integração é finito.

Seja $y = f(x)$ uma função cujos valores $f(x_0, f(x_1), \ldots, f(x_n)$ são conhecidos (por exemplo, por meio de uma tabela). De (11.2), temos que:

$$\int_a^b f(x)\ dx \ = \int_{x_0}^{x_n} f(x)\ dx \ \cong \ \sum_{k=0}^{n} f_k \int_{x_0}^{x_n} \ell_k(x)\ dx.$$

Supondo então os argumentos x_i igualmente espaçados de h e considerando $u = \dfrac{x - x_0}{h}$; temos que:

$$dx = h\ du \quad \text{e quando} \quad \begin{array}{l} x = x_0 \ \rightarrow \ u = 0\ , \\ x = x_n \ \rightarrow \ u = n\ . \end{array}$$

Logo:

$$\int_{x_0}^{x_n} f(x)\ dx \cong \sum_{k=0}^{n} f_k\ h \int_0^n \lambda_k(u)\ du.$$

onde os λ_k são os polinômios usados na fórmula de Lagrange para argumentos igualmente espaçados. Fazendo:

$$\int_0^n \lambda_k(u)\ du \ = \ C_k^n\ ; \tag{11.3}$$

obtemos:

$$\int_{x_0}^{x_n} f(x)\ dx \ \cong \ \sum_{k=0}^{n} f_k\ h\ C_k^n\ . \tag{11.4}$$

Observe que a fórmula (11.4) independe dos limites de integração. Trataremos de obter, agora, algumas fórmulas de Newton-Cotes. Mais adiante analisaremos o termo do resto.

1º Caso: Consideremos $\mathbf{n = 1}$; isto é, queremos obter uma fórmula para integrar $f(x)$ entre dois pontos consecutivos x_0 e x_1, usando polinômio do primeiro grau. Temos, em vista de (11.4), que:

$$\int_{x_0}^{x_1} f(x)dx \ \cong \ \sum_{k=0}^{1} f_k\ h\ C_k^1\ ; \tag{11.5}$$

onde, de (11.3),

$$\begin{aligned} C_0^1 \ &= \ \int_0^1 \lambda_0(u)du \ = \ \int_0^1 \frac{u-1}{0-1}\ du \ = \ \int_0^1 (1-u)du \ = \ \frac{1}{2}\ , \\ C_1^1 \ &= \ \int_0^1 \lambda_1(u)du \ = \ \int_0^1 \frac{u-0}{1-0}\ du \ = \ \frac{1}{2}\ . \end{aligned}$$

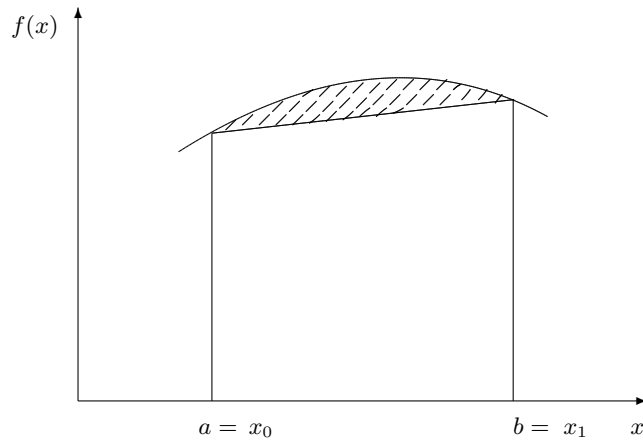
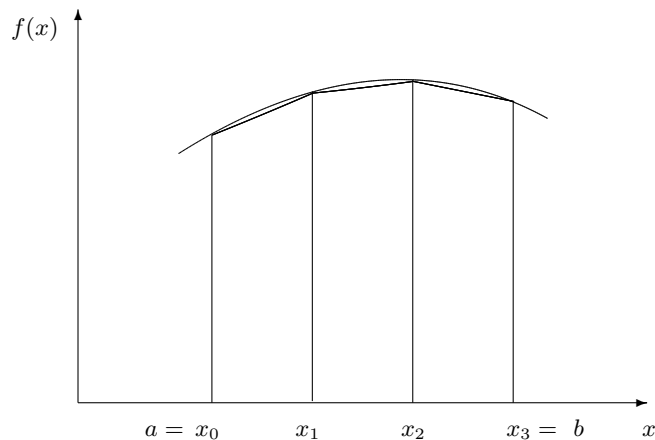
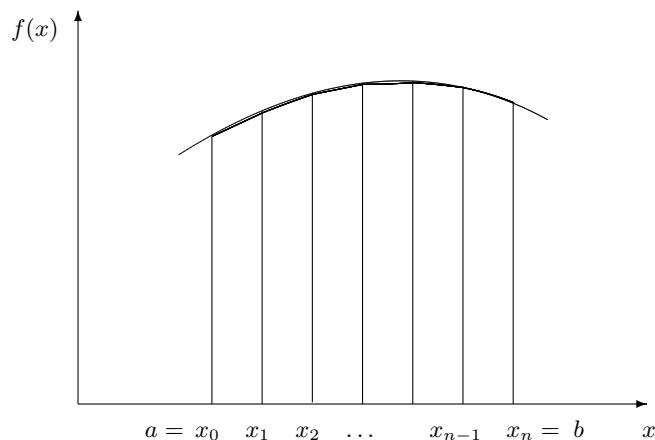
Portanto, substituindo C_0^1 e C_1^1 em (11.5), obtemos:

$$\int_{x_0}^{x_1} f(x)\, dx \;\cong\; h\;\left[\frac{1}{2} f\left(x_0\right) \;+\; \frac{1}{2} f\left(x_1\right)\right] \;,$$

ou

$$\int_{x_0}^{x_1} f(x)\, dx \;\cong\; \frac{h}{2}\;\left[f\left(x_0\right) \;+\; f\left(x_1\right)\right] \;, \tag{11.6}$$

que é uma fórmula de Newton-Cotes, conhecida como **Regra do Trapézio**.

Observe que: se o intervalo $[a, b]$ é pequeno a aproximação é razoável; mas se $[a, b]$ é grande o erro também pode ser grande. Na figura a seguir, a área hachurada é o erro cometido ao calcularmos a integral de a até b, usando (11.6).

Assim, se o intervalo de integração é grande podemos dividir o intervalo $[a, b]$ em N sub-intervalos de amplitude $h = \dfrac{b-a}{N}$ de tal forma que $x_0 = a$; $x_N = b$ e em cada sub-intervalo $[x_j, x_{j+1}]$, $j = 0, 1, \ldots, N-1$, aplicar a Regra do Trapézio. O erro agora é a soma das áreas entre a curva e as retas, como mostrado na figura a seguir:

Observe ainda que, quando $h \to 0$, estaremos tendendo ao resultado exato da integral pois o erro estará tendendo a zero, como pode ser visto na seguinte figura:

Assim, utilizando o que foi descrito, obtemos:

$$\begin{aligned}\int_{x_0}^{x_N} f(x)dx &= \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx \\ &+ \ldots + \int_{x_{N-1}}^{x_N} f(x)dx \\ &\simeq \frac{h}{2}\left[f(x_0) + f(x_1)\right] + \frac{h}{2}\left[f(x_1) + f(x_2)\right] \\ &+ \ldots + \frac{h}{2}\left[f(x_{N-1}) + f(x_N)\right] .\end{aligned}$$

Na expressão acima vemos que com exceção da f calculada nos pontos x_0 e x_N, todas as demais aparecem duas vezes. Portanto, podemos escrever:

$$\int_{x_0}^{x_N} f(x)dx = \frac{h}{2}\left[f(x_0) + 2\left(f(x_1) + f(x_2) + \ldots + f(x_{N-1})\right) + f(x_N)\right] , \tag{11.7}$$

e assim obtemos a **Regra do Trapézio Generalizada**. Na prática, só utilizamos esta regra.

Exemplo 11.2 - *Calcular, usando a regra do trapézio,*

$$\int_0^{1.2} e^x \cos x \, dx .$$

Solução: Temos que:

	x_0	x_1	x_2	x_3	x_4	x_5	x_6
x	0	0.2	0.4	0.6	0.8	1.0	1.2
e^x	1	1.221	1.492	1.822	2.226	2.718	3.320
$\cos x$	1	0.980	0.921	0.825	0.697	0.540	0.362
$e^x \cos x$	1	1.197	1.374	1.503	1.552	1.468	1.202

Aplicando, (11.7), obtemos:

$$\begin{aligned}
\int_0^{1.2} e^x \cos x \, dx &= \frac{h}{2} \left[f(x_0) + 2\left(f(x_1) + \ldots + f(x_5)\right) + f(x_6) \right] \\
&= \frac{0.2}{2} [1 + 2\ (1.197 + 1.374 + 1.503 + 1.552 + 1.468) + 1.202] \\
&= 0.1[1 + 2(7.094) + 1.202] \\
&= 0.1[1 + 14.188 + 1.202] \ = \ 0.1[16.39] \ = \ 1.639.
\end{aligned}$$

Exercícios

11.1 - *Aplicar a regra do trapézio para calcular:*

$$\int_{1.00}^{1.30} \sqrt{x}\, dx \ ,$$

utilizando os dados da tabela a seguir:

x	1.00	1.05	1.10	1.15	1.20	1.25	1.30
$\sqrt{x}$	1.0000	1.0247	1.0488	1.0723	1.0954	1.1180	1.1401

11.2 - *Calcular:*

$$\int_0^{0.8} \cos x \, dx \ ,$$

pela regra do trapézio, com $h = 0.4,\ 0.2,\ e\ 0.1$, *sabendo que:*

x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\cos x$	1	0.995	0.980	0.955	0.921	0.877	0.825	0.764	0.696

11.3 - *Usando a regra do trapézio sobre 5 pontos, calcular:*

$$\int_{1.2}^{1.6} sen\ x \, dx \ .$$

Sabe-se que:

x	1.2	1.3	1.4	1.5	1.6
$sen\ x$	0.93204	0.96356	0.98545	0.99749	0.99957

11.4 - *Dada a tabela:*

x	0	0.2	0.4	0.6	0.8
e^x	1	1.22	1.49	1.82	2.22

calcular:

$$\int_0^{0.8} x\, e^x \, dx \ ,$$

pela regra do trapézio usando todos os pontos.

11.5 - *Provar que:* $C_k^n = C_{n-k}^n$.

Sugestão: Faça uma mudança de variável.

2º caso: Consideremos $\mathbf{n = 2}$; isto é, queremos obter uma fórmula para integrar $f(x)$ entre três pontos consecutivos x_0, x_1 e x_2 usando polinômio do 2º grau. Temos, de (11.4), que:

$$\int_{x_0}^{x_2} f(x)\, dx \simeq \sum_{k=0}^{2} f_k\, h\, C_k^2 , \tag{11.8}$$

onde, de (11.3),

$$\begin{aligned} C_0^2 &= \int_0^2 \lambda_0(u)\, du = \int_0^2 \frac{(u-1)(u-2)}{(0-1)(0-2)}\, du \\ &= \frac{1}{2} \int_0^2 \left(u^2 - 3u + 2\right)\, du = \frac{1}{3} , \\ C_1^2 &= \int_0^2 \lambda_1(u)\, du = \int_0^2 \frac{(u-0)(u-2)}{(1-0)(1-2)}\, du \\ &= - \int_0^2 \left(u^2 - 2u\right)\, du = \frac{4}{3} , \end{aligned}$$

e pelo exercício 11.5 temos que: $C_2^2 = C_0^2 = \frac{1}{3}$.

Substituindo os valores de $C_k^2,\ k = 0, 1, 2$ em (11.8), obtemos:

$$\int_{x_0}^{x_2} f(x)dx \simeq h\left[\frac{1}{3} f\left(x_0\right) + \frac{4}{3} f\left(x_1\right) + \frac{1}{3} f\left(x_2\right)\right] ,$$

ou

$$\int_{x_0}^{x_2} f(x)dx \simeq \frac{h}{3}\left[f\left(x_0\right) + 4 f\left(x_1\right) + f\left(x_2\right)\right] , \tag{11.9}$$

que é uma fórmula de Newton-Cotes, conhecida como **Regra $\frac{1}{3}$ de Simpson**.

De maneira análoga à regra do trapézio, a generalização da regra $\frac{1}{3}$ de Simpson para integração ao longo de um intervalo $[a, b]$, é feita dividindo-se $[a, b]$ num número **par** $2N$ de sub-intervalos de amplitude $h = \frac{b-a}{2N}$ de tal forma que $x_0 = a,\ x_{2N} = b$. Note que o número de subdivisões deve ser múltiplo de 2, pois precisamos de dois subintervalos (e portanto de três pontos) para aplicar uma vez a regra (11.9). Temos então:

$$\begin{aligned} \int_{x_0}^{x_{2N}} f(x)\, dx &= \int_{x_0}^{x_2} f(x)\, dx + \int_{x_2}^{x_4} f(x)dx \\ &+ \ldots + \int_{x_{2N-2}}^{x_{2N}} f(x)dx . \end{aligned}$$

Usando a regra $\frac{1}{3}$ de Simpson ao longo do intervalo $[x_j, x_{j+2}], j = 0, 2, \ldots, 2N-2$, temos:

$$\begin{aligned}\int_{x_0}^{x_{2N}} f(x)\, dx &\simeq \frac{h}{3}\left[f(x_0) + 4f(x_1) + f(x_2)\right] \\ &+ \frac{h}{3}\left[f(x_2) + 4f(x_3) + f(x_4)\right] \\ &+ \ldots + \frac{h}{3}\left[f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\right] .\end{aligned}$$

Na expressão acima vemos que com exceção da f calculada nos pontos x_0 e x_{2N}, as f calculadas nos pontos extremos do intervalo de integração aparecem duas vezes. Portanto, podemos escrever:

$$\begin{aligned}\int_{x_0}^{x_{2N}} f(x)dx &\simeq \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 2f(x_2) + 4f(x_3) + 2f(x_4)\right. \\ &+ \left.\ldots + 2f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\right] ,\end{aligned} \tag{11.10}$$

e assim obtemos a **Regra $\frac{1}{3}$ de Simpson Generalizada**. Novamente, na prática, só utilizamos esta regra.

Exemplo 11.3 - *Usando a regra $\frac{1}{3}$ de Simpson, calcular a integral do exemplo 11.2.*

Solução: Temos, usando (11.11), que:

$$\begin{aligned}\int_0^{1.2} e^x \cos x\, dx &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3)\right. \\ &+ \left.2f(x_4) + 4f(x_5) + f(x_6)\right] \\ &= \frac{0.2}{3}\left[1 + 4(1.197 + 1.503 + 1.468) +\right. \\ &+ \left.2(1.374 + 1.552) + 1.202\right] \\ &= \frac{0.2}{3}\left[1 + 4(4.168) + 2(2.926) + 1.202\right] \\ &= \frac{0.2}{3}\left[1 + 16.672 + 5.852 + 1.202\right] \\ &= \frac{0.2}{3}[24.726] = 1.6484 .\end{aligned}$$

Exercícios

11.6 - *Resolver os exercícios 11.1, 11.2, 11.3 e 11.4 usando a Regra $\frac{1}{3}$ de Simpson.*

11.7 - *A velocidade v de um foguete lançado do chão verticalmente (para cima, é claro) foi tabelada como se segue:*

$t(seg)$	0	5	10	15	20
$v(pés/seg)$	0	60.6	180.1	341.6	528. 4

Usando a regra $\frac{1}{3}$ *de Simpson calcular a altura do foguete após 20 segundos.*

11.8 - *Usando a regra* $\frac{1}{3}$ *de Simpson, calcular:*

$$\int_{1.0}^{1.6} \ln\ x\ dx\ .$$

Sabe-se que:

x	1.0	1.1	1.2	1.3	1.4	1.5	1.6
$\ln\ n\ x$	0	0.095	0.182	0.262	0.336	0.405	0.470

3º Caso: Consideremos $n = 3$, isto é, queremos obter uma fórmula para integrar $f(x)$ entre 4 pontos consecutivos x_0, x_1, x_2, x_3, usando polinômio do 3º grau. Temos, em vista de (11.4), que:

$$\int_{x_0}^{x_3} f(x)\ dx\ \simeq\ \sum_{k=0}^{3} f_k\ h\ C_k^3, \tag{11.11}$$

onde, de (11.3),

$$\begin{aligned} C_0^3 &= \int_0^3 \lambda_0(u)du\ =\ \int_0^3\ \frac{(u-1)(u-2)(u-3)}{(0-1)(0-2)(0-3)}\ du \\ &= -\ \frac{1}{6}\ \int_0^3\ \left(u^3 - 6u^2 + 11u - 6\right)\ du\ =\ \frac{3}{8}\ , \\ C_1^3 &= \int_0^3 \lambda_1(u)du\ =\ \int_0^3\ \frac{u(u-2)(u-3)}{(1-0)(1-2)1-3)}\ du \\ &= \frac{1}{2}\ \int_0^3\ \left(u^3 - 5u^2 + 6u\right)\ du\ =\ \frac{9}{8}\ . \end{aligned}$$

Pelo exercício 11.5, temos que: $C_3^3 = C_0^3 = \frac{3}{8};\ C_2^3 = C_1^3 = \frac{9}{8}$.

Assim, substituindo os valores de $C_k^3,\ k = 0, 1, 2, 3$, em (11.11), obtemos:

$$\int_{x_0}^{x_3} f(x)\ dx\ \simeq\ h\left[\frac{3}{8}f(x_0) + \frac{9}{8}f(x_1) + \frac{9}{8}f(x_2) + \frac{3}{8}f(x_3)\right]\ ,$$

ou

$$\int_{x_0}^{x_3} f(x)\ dx\ \simeq\ \frac{3}{8}h\left[f(x_0) + 3\left(f(x_1) + f(x_2)\right) + f(x_3)\right]\ , \tag{11.12}$$

que é uma fórmula de Newton-Cotes conhecida como **Regra $\frac{3}{8}$ de Simpson**.

Para generalizar a regra $\frac{3}{8}$ de Simpson devemos dividir o intervalo $[a,b]$ em um número de subintervalos de amplitude $h = \frac{b-a}{3N}$ de tal forma que $x_0 = a, \ x_{3N} = b$. Note que o número de subdivisões deve ser múltiplo de 3, pois precisamos de três subintervalos (e portanto de quatro pontos) para aplicar uma vez a regra (11.12). Temos então:

$$\begin{aligned} \int_{x_0}^{x_{3N}} f(x)\ dx &= \int_{x_0}^{x_3} f(x)\ dx + \int_{x_3}^{x_6} f(x)\ dx \\ &+ \ \ldots + \int_{x_{3N-3}}^{x_{3N}} f(x)\ dx \end{aligned}$$

Usando a regra $\frac{3}{8}$ de Simpson ao longo do intervalo $[x_j, x_{j+3}]$, $j = 0, 3, 6, \ldots, 3N-3$, obtemos:

$$\begin{aligned} \int_{x_0}^{x_{3N}} f(x)\ dx &\simeq \frac{3}{8}h\ \left[f(x_0) + 3\left(f(x_1\ ) + f)\, x_2\ \right)) + f(x_3)\right] \\ &+ \frac{3}{8}h\left[f(x_3) + 3(f x_4) + f(x_5)\ ) + f(x_6)\right] \\ &+ \ \ldots\ldots \\ &+ \frac{3}{8}h\left[f(x_{3N-3}) + 3\left((x_{3N-2}) + f(x_{3N-1})\right) + f(x_{3N})\right]\ . \end{aligned}$$

Novamente, na expressão acima vemos que com exceção da f calculada nos pontos x_0 e x_{3N}, as demais calculadas nos pontos extremos do intervalo de integração aparecem duas vezes. Portanto, podemos escrever:

$$\begin{aligned} \int_{x_0}^{x_{3N}} f(x)dx &\simeq \frac{3}{8}h\left[f(x_0) + 3\left(f(x_1) + f(x_2)\right) + 2f(x_3)\ + \right. \\ &+ \ 3\left(f(x_4) + f(x_5)\right) + 2f(x_6)\ + \ldots\ldots \\ &+ \ \left. 2f(x_{3N-3}) + 3\left(f(x_{3N-2}) + f(x_{3N-1})\right) + f(x_{3N})\ \right]\ . \end{aligned} \tag{11.13}$$

e assim obtemos a **Regra $\frac{3}{8}$ de Simpson Generalizada**. Novamente, na prática, só utilizamos esta regra.

Exemplo 11.4 - *Usando a regra $\frac{3}{8}$ de Simpson, calcular a integral do exemplo 11.2.*

Solução: Usando (11.13), temos que:

$$\begin{aligned}
\int_0^{1.2} e^x \ \cos \ x \ dx \ &\cong \ \frac{3}{8}h\,[\ f\,(x_0) + 3(f\,(x_1) + f\,(x_2)) + 2f\,(x_3) \\
&+ \ 3(f\,(x_4) + f\,(x_5)) + f\,(x_6)\ ] \\
&= \ \frac{0.6}{8}\ [\ 1 + 3(1.197 + 1.374) + 2(1.503) \\
&+ \ 3(1.552 + 1.468) + 1.202\ ] \\
&= \ \frac{0.6}{8}\,[1 + 3(2.571) + 3.006 + 3(3.020) + 1.202] \\
&= \ \frac{0.6}{8}\,[1 + 7.713 + 3.006 + 9.060 + 1.202] \\
&= \ \frac{0.6}{8}[21.981] = 1.648575\ .
\end{aligned}$$

Observações:

1) De maneira semelhante ao descrito nesta seção, podemos obter outras fórmulas de Newton-Cotes do tipo fechado. Para tanto, basta irmos aumentando o grau do polinômio de interpolação. Entretanto, na prática, as fórmulas mais usadas são a regra do Trapézio e as de Simpson, pois quando $h \to 0$, todas tendem ao resultado exato da integral.

2) Calculando diretamente a integral, obtemos:

$$\int_0^{1.2} e^x \ \cos \ x \ dx \ = \ \frac{e^x}{2}(sen\ x + \cos\ x)\Big]_0^{1.2} \ ,$$

e usando 6 casas decimais, (trabalhando com arredondamento), segue que o resultado da integral é **1.6487747**. Assim, se compararmos os resultados obtidos nos exemplos anteriores vemos que temos uma casa decimal precisa quando utilizamos a regra do trapézio e três casas decimais precisas nas regras de Simpson. Veremos a seguir, quando estudarmos o termo do resto para as fórmulas de Newton-Cotes, que ambas as regras de Simpson possuem a mesma ordem de convergência e portanto tanto faz utilizar uma como a outra que teremos o resultado da integral com o mesmo número de casas decimais corretas. Uma pergunta que surge naturalmente é: porquê então duas regras? A resposta está no fato que a aplicação das regras dependem das subdivisões dos intervalos. Note que na prática podemos utilizar as duas regras de Simpson para resolver uma integral, como é mostrado no exemplo a seguir.

Exemplo 11.5 - *Usando a regra de Simpson, calcular a integral de 0 a 1.0 da função dada no exemplo 11.2.*

Solução: Observe que neste caso não podemos utilizar apenas uma das regras de Simpson para resolver a integral, visto que o número de pontos, para ambas, não é adequado . Podemos entretanto utilizar a regra $\frac{1}{3}$ de Simpson para integrar de 0 a 0.4 e a regra $\frac{3}{8}$ para integrar de 0.4 a 1.0 ou então utilizar a

regra $\frac{3}{8}$ de Simpson no intervalo $[0, 0.6]$ e a regra $\frac{1}{3}$ de Simpson no intervalo $[0.6, 1.0]$. Consideremos a primeira possiblidade. Temos então:

$$\begin{aligned}
\int_0^{1.0} e^x \ \cos \ x \ dx &\cong \frac{1}{3}h\,[\ f\,(x_0) + 4f\,(x_1) + f\,(x_2)\ ] \\
&+ \frac{3}{8}h\,[\ f\,(x_2) + 3(f\,(x_3) + f\,(x_4)) + f\,(x_5)\ ] \\
&= \frac{0.2}{3}\ [\ 1 + 4(1.197) + 1.374\ ] \\
&+ \frac{0.6}{8}\ [\ 1.374 + 3(1.503 + 1.552) + 1.468\ ] \\
&= \frac{0.2}{3}[1 + 4.788 + 1.374] \\
&+ \frac{0.6}{8}\ [\ 1.374 + 9.165 + 1.468\ ] \\
&= \frac{0.2}{3}(7.162) + \frac{0.6}{8}(12.007) \\
&= 0.477467 + 0.900525 = 1.377992\ .
\end{aligned}$$

Observe que não podemos aplicar as regras de Simpson com a regra do trapézio, pois esta última possui ordem de convergência menor, como veremos na próxima seção.

Exercícios

11.9 - *Resolver os exercícios 11.1 e 11.8 usando a regra* $\frac{3}{8}$ *de Simpson.*

11.10 - *Calcular:*

$$\int_0^{0.6} \cos x\ dx\ ,$$

pela regra $\frac{3}{8}$ *de Simpson; com* $h = 0.1$*. Use a tabela do exercício 11.2.*

11.11 - *Usando a regra* $\frac{3}{8}$ *de Simpson e* $h = 0.6,\ 0.2$*, calcular :*

$$\int_0^{1.2} e^{-x}\ sen\ x\ dx\ .$$

Sabe-se que:

x	0	0.2	0.4	0.6	0.8	1.0	1.2
e^{-x}	1.000	0.819	0.670	0.548	0.449	0.367	0.301
$sen\ x$	0	0.198	0.398	0.565	0.717	0.841	0.932

11.2.2 Erro nas Fórmulas de Newton-Cotes

Estudaremos nesta seção o termo do resto ou o erro que cometemos ao aproximar o valor de uma integral usando as fórmulas de Newton-Cotes do tipo fechado, isto é, estudaremos o termo $R(f)$ dado em (11.1). Para tanto enunciaremos dois teoremas, cujas demonstrações aqui omitidas, podem ser encontradas, por exemplo em [Jennings, 19..], e cujos resultados são extremamente importantes.

Teorema 11.2 - *Se os pontos $x_j = x_0 + jh, \ j = 0, 1, \ldots, n$ dividem $[a, b]$ em um número* **ímpar** *de intervalos iguais e $f(x)$ tem derivada de ordem $(n+1)$ contínua em $[a, b]$, então a expressão do erro para as fórmulas de Newton-Cotes do tipo fechado, com n ímpar, é dada por:*

$$R(f) = \frac{h^{n+2} f^{(n+1)}(\xi)}{(n+1)!} \int_0^n u(u-1)\ldots(u-n)\, du \ ,$$

para algum ponto $\xi \in [a, b]$.

Teorema 11.3 - *Se os pontos $x_j = x_0 + jh, \ j = 0, 1, \ldots, n$ dividem $[a, b]$ em um número* **par** *de intervalos iguais e $f(x)$ tem derivada de ordem $(n+2)$ contínua em $[a, b]$, então a expressão do erro para as fórmulas de Newton-Cotes do tipo fechado, com n par, é dada por:*

$$R(f) = \frac{h^{n+3} f^{(n+2)}(\xi)}{(n+2)!} \int_0^n \left(u - \frac{n}{2}\right) u(u-1)\ldots(u-n)\, du \ ,$$

para algum ponto $\xi \in [a, b]$.

Assim, o erro da fórmula trapezoidal sobre o intervalo $[x_0, x_1]$ é obtido do Teorema 11.2, colocando $\mathbf{n = 1}$, isto é:

$$\begin{aligned} R(f) &= \frac{h^3 f''(\xi)}{2!} \int_0^1 [u(u-1)]\, du \\ &= \frac{h^3 f''(\xi)}{2!} \int_0^1 [(u^2 - u)]\, du \\ &= \frac{h^3 f''(\xi)}{2!} \left(-\frac{1}{6}\right) \end{aligned}$$

Portanto o erro, ao aplicarmos uma vez a regra do trapézio, é dado por:

$$R(f) = -\frac{h^3 f''(\xi)}{12} \quad , \quad x_0 < \xi < x_1 \ . \tag{11.14}$$

Assim, podemos escrever:

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2} \left[h(x_0) + f(x_1)\right] - \frac{h^3}{12} f''(\xi) \ ; \ x_0 < \xi < x_1 \ .$$

O erro na fórmula do trapézio generalizada é obtido adicionando-se N erros da forma (11.14); onde $N = \frac{b-a}{h}$. Logo:

$$\begin{aligned}
\int_{x_0}^{x_N} f(x)dx &= [f(x_0) + 2(f(x_1) + \ldots + f(x_{N-1})) + f(x_N)] \\
&- \frac{Nh^3}{12} f''(\xi) \quad , \quad x_0 < \xi < x_N \\
&= \frac{h}{2} [f(x_0) + 2(f(x_1) + \ldots + f(x_{N-1})) + f(x_N)] \\
&- \frac{(b-a)}{12} h^2 f''(\xi) \ , \quad x_0 < \xi < x_N \ .
\end{aligned}$$

Observe que o termo do resto não deve, na prática, ser subtraído do resultado aproximado, obtido pela aplicação da regra do trapézio, visto que nunca conseguíriamos o resultado exato, pois o ponto ξ que fornece a igualdade, existe e é único mas não temos como determiná-lo. Assim a aplicação da fórmula do termo do resto é útil quando queremos o resultado com uma precisão pré-fixada, como mostraremos no exemplo a seguir.

Exemplo 11.6 - *Determinar o menor número de intervalos em que podemos dividir [0,1.2] para obter:*

$$\int_0^{1.2} e^x \cos x \, dx \ ,$$

pela regra do trapézio com três casas decimais precisas.

Solução: Como queremos determinar o menor número de intervalos, devemos usar a expressão do erro para a fórmula do trapézio generalizada. Assim:

$$R(f) = \frac{Nh^3}{12} f''(\xi), \ x_0 < \xi < x_N \ ,$$

$$\Rightarrow |R(f)| \leq \frac{Nh^3}{12} \max_{0 \leq t \leq 1.2} |f''(t)| \ , \tag{11.15}$$

desde que não podemos determinar $f''(\xi)$. Devemos então calcular a derivada segunda da f e avaliar seu valor máximo no intervalo $[0, 1.2]$. Temos:

$$\begin{aligned}
f(t) &= e^t \cos t \quad , \quad f'(t) = e^t(\cos t - sen\ t) \ , \\
f''(t) &= e^t(\cos t - sen\ t) - e^t(sen\ t + \cos t) \\
\Rightarrow \quad & f''(t) = -2 e^t \ sen\ t \ .
\end{aligned}$$

Portanto:

$$\max_{0 \leq t \leq 1.2} |f''(t)| \; = \; |f(1.2)| \; = \; 2(3.320)(0.932) \; = \; 6.188 \; .$$

Substituindo o valor máximo da derivada segunda em (11.15), lembrando que $h = \dfrac{b-a}{N} = \dfrac{1.2-0}{N} = \dfrac{1.2}{N}$, e impondo que o erro seja $\leq 0.5 \times 10^{-3}$, pois queremos o resultado com três casas decimais corretas, obtemos:

$$\begin{array}{lcl} R(f) & \leq & \dfrac{1.2\; h^2}{12}(6.188) \; \leq \; 0.5 \; \times \; 10^{-3} \\ \Rightarrow & & h^2 \; \leq \; 0.0000808 \\ \Rightarrow & & h \; \leq \; 0.02842 \; . \end{array}$$

Observe que devemos escolher o maior h, que seja ≤ 0.02842, mas que divida **exatamente** o intervalo $[0, 1.2]$. Assim, escolhemos $h = 0.025$, e usando o fato que $N = \dfrac{b-a}{h}$, obtemos que $N_{min} = 48$. Assim devemos dividir o intervalo [0,1.2] em 48 sub-intervalos iguais para obter $\int_0^{1.2} e^x \cos x \, dx$ pela regra do trapézio com três casas decimais precisas. Compare com as regras de Simpson, onde a mesma precisão é obtida com um número muito menor de subintervalos.

Para obtermos o erro na fórmula $\frac{1}{3}$ de Simpson, sobre o intervalo $[x_0, x_2]$, substituimos **n** por 2 no Teorema 11.3. Assim:

$$\begin{array}{rcl} R(f) & = & \dfrac{h^5 f^{(IV)}(\xi)}{4!} \displaystyle\int_0^2 [(u-1)u(u-1)(u-2)]\; du \\ & = & \dfrac{h^5 f^{(IV)}(\xi)}{24} \displaystyle\int_0^2 \left(u^4 - 4u^3 + 5u^2 - 2u\right)\; du \\ & = & \dfrac{h^5 f^{(IV)}(\xi)}{4!} \left(-\dfrac{4}{15}\right) \; . \end{array}$$

Portanto o erro, ao aplicarmos uma vez a regra $\frac{1}{3}$ de Simpson, é dado por:

$$R(f) \; = \; -\frac{h^5}{90} \; f^{(IV)}(\xi) \quad , \quad x_0 < \xi < x_2 \; . \tag{11.16}$$

Então, podemos escrever:

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3} \; [f(x_0) + 4f(x_1) + f(x_2)] \; - \; \frac{h^5}{90} f^{(IV)}(\xi), \; x_0 < \xi < x_2 \; .$$

O erro na fórmula $\frac{1}{3}$ de Simpson generalizada é obtida adicionando-se N erros da forma (11.16) onde $N = \frac{b-a}{2h}$. Assim:

$$\begin{aligned}
\int_{x_0}^{x_{2N}} f(x)dx &= \frac{h}{3} [\ f(x_0) + 4f(x_1) + 2f(x_2) \\
&+ \ \ldots + 2f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\ ] \\
&- \ \frac{Nh^5}{90}\ f^{(IV)}(\xi)\ , x_0 < \xi < x_{2N}. \\
&= \frac{h}{3} [\ f(x_0) + 4f(x_1) + 2f(x_2) \\
&+ \ \ldots + 2f(x_{2N-2}) + 4f(x_{2N-1}) + f(x_{2N})\ ] \\
&- \ \frac{(b-a)h^4}{180}\ f^{(IV)}(\xi)\ , x_0 < \xi < x_{2N}.
\end{aligned}$$

Deixamos como exercício verificar que a expressão do erro para a regra $\frac{3}{8}$ de Simpson, e para regra $\frac{3}{8}$ de Simpson generalizada, são dadas, respectivamente, por: é dada por:

$$R(f) \;=\; -\frac{3}{80}\ h^5\ f^{(IV)}(\xi) \quad , \quad x_0 < \xi < x_3\ ,$$

$$\frac{(b-a)h^4}{80}\ f^{(IV)}(\xi)\ , x_0 < \xi < x_{3N}\ ,$$

desde que $N = \frac{(b-a)}{3h}$.

Observações:

1) Pelas expressões do erro, vemos que as fórmulas de Simpson são da ordem de h^4, em símbolo, $O(h^4)$, enquanto que a regra do trapézio é da $O(h^2)$. Assim, as regras de Simpsom possuem a mesma ordem de convergência, portanto ambas convergem para o resultado exato com a mesma velocidade. Para exemplificar o significado da importância da ordem de convergência, consideremos que a aplicação das regras de Simpson com um determinado h fornecem o resultado com um erro inferior a 10^{-3}. A aplicação da mesma regra com $\bar{h} = \frac{h}{10}$ fornecerá o resultado com erro inferior a 10^{-7}. Por outro lado se o resultado da aplicação da regra do trapézio tiver erro inferior a 10^{-1}, com um determinado h, então o erro será inferior a 10^{-3} com $\bar{h} = \frac{h}{}$. Portanto quando $h \to 0$, as regras de Simpson convergem mais rapidamente para o resultado exato da integral.

2) Apesar da fórmula $\frac{1}{3}$ de Simpson ter sido obtida aproximando-se a função por polinômio de grau 2, ela é exata também para polinômios de grau 3, visto que na fórmula do erro aparece a derivada quarta da função f. Pode ser demonstrado que se n é **par** então as fórmulas de Newton-Cotes do tipo fechado têm **grau de precisão** $n+1$.

3) Para obter o resultado de uma integral com uma determinada precisão, podemos utilizar a fórmula do erro, impondo que a mesma, em módulo, seja inferior a 0.5×10^{-k}, onde k é o número de casas decimais corretas que desejamos no resultado, e assim obter o número de intervalos necessários, (ver

exemplo 11,5), **ou** ir aumentando o número de pontos e comparando dois resultados consecutivos, até obter a precisão desejada, como é mostrado no exemplo a seguir. Na prática usamos esta segunda possiblidade.

Exemplo 11.7 - *Usando a regra $\frac{1}{3}$ de Simpson, obter a integral do exemplo 11.2, com 2 casas decimais corretas.*

Solução: Inicialmente calculamos a integral usando apenas 3 pontos. Assim:

$$\begin{aligned}
\int_0^{1.2} e^x \ \cos \ x \ dx &\cong \frac{1}{3}h\,[\ f\,(0) + 4f\,(0.6) + f\,(1.2)\ ] \\
&= \frac{1}{3}(0.6)(1 + 4(1.503) + 1.202) \\
&= 1.6428 = I_3
\end{aligned}$$

Agora calculamos a integral com 5 pontos. Assim:

$$\begin{aligned}
\int_0^{1.2} e^x \ \cos \ x \ dx &\cong \frac{1}{3}h\,[\ f\,(0) + 4f\,(0.3) + 2f\,(0.6) + 4f\,(0.9) + (1.2)\ ] \\
&= \frac{1}{3}(0.3)(1 + 4(1.289) + 2(1.503) + 4(1.530) + 1.202) \\
&= 1.6464 = I_5
\end{aligned}$$

desde que $f(0.3) = 1.289$ e $f(0.9) = 1.530$. Comparando $I_5 = 1.6464$ com $I_3 = 1.6428$ vemos que as duas primeiras casas decimais permanecem inalteradas. Portanto o valor da integral com duas casas decimais corretas é 1.64.

Observe que o exemplo dado ilustra o procedimento a ser adotado, no computador devemos tomar $h \to 0$, isto é: $h = 0.1, 0.01, \ldots$ e ir comparando os resultados.

Exercícios

11.12 - *Determine h de modo que a regra do trapézio forneça o valor de:*

$$I \ = \ \int_0^1 \ e^{-x^2} \ dx \ ,$$

com erro inferior a $0.5 \ \times \ 10^{-6}$.

11.13 - *Achar o número mínimo de intervalos que se pode usar para, utilizando a regra $\frac{1}{3}$ de Simpson, obter*

$$\int_0^{\pi/2} \ e^{-x} \ \cos \ x \ dx \ ,$$

com 4 casas decimais corretas.

11.14 - *Determinar h de modo que a regra $\frac{3}{8}$ de Simpson forneça o valor de*

$$\int_{0.2}^{0.8} sen\ x\ dx \ ,$$

com erro inferior a 0.5×10^{-3}.

11.3 Polinômios Ortogonais

Ao lado das fórmulas de Newton-Cotes para integração numérica, as fórmulas de quadratura de Gauss, a serem definidas mais adiante, se destacam por fornecerem resultados altamente precisos. Tais fórmulas, baseiam-se em propriedades de polinômios ortogonais os quais passamos a estudar agora. (Para melhor entendimento das propriedades reveja os conceitos de base ortogonal, mudança de base e produto escalar, dados no Capítulo 1).

Sejam $\phi_0(x), \phi_1(x), \phi_2(x), \ldots$, uma família de polinômios de graus $0, 1, 2, \ldots$.

Se:

$$\begin{cases} & (\phi_i(x), \phi_j(x)) \ = \ 0 \quad , \quad \text{para} \quad i \neq j \ , \\ \text{e} & \\ & (\phi_i(x), \phi_i(x)) \ \neq \ 0 \quad , \quad \text{para} \quad \phi_i \neq 0 \ , \end{cases} \tag{11.17}$$

então os polinômios $\phi_0(x), \phi_1(x), \phi_2(x), \ldots$ se dizem ortogonais.

Neste estudo, estamos considerando o produto escalar:

$$(f, g) \ = \ \int_a^b \omega(x)\ f(x)\ g(x)\ dx \quad , \tag{11.18}$$

com $\omega(x) \geq 0$ e contínua em $[a, b]$, onde $\omega(x)$ é a função peso.

Os polinômios $\phi_i(x),\ i = 0, 1, 2, \ldots$, podem ser obtidos pela ortogonalização da sequência $\{1, x, x^2, \ldots\}$ (usando o processo de ortogonalização de Gram-Schmidt (Capítulo 1)) ou, através do seguinte:

Teorema 11.4 - *Sejam os polinômios* $\phi_0(x), \phi_1(x), \phi_2(x), \ldots,$ *de graus* $0, 1, 2, \ldots,$ *definidos por:*

$$\begin{cases} \phi_0(x) \ = \ 1 \ , \\ \phi_1(x) \ = \ x \ - \ \dfrac{(x\ \phi_0(0), \phi_0(0))}{(\phi_0(0), \phi_0(x))}\ \phi_0(x) \ = \ x \ - \ \dfrac{(x, 1)}{(1, 1)}\ 1 \ , \\ \text{e, para } k = 1, 2, 3, \ldots, \\ \phi_{k+1}(x) \ = \ x\ \phi_k(x) - \alpha_k \phi_k(x) \ - \ \beta_k\ \phi_{k-1}(x) \ , \end{cases} \tag{11.19}$$

onde:

$$\alpha_k \ = \ \frac{(x\ \phi_k(x),\ \phi_k(x))}{(\phi_k(x), \phi_k(x))} \quad ; \quad \beta_k \ = \ \frac{(\phi_k(x),\ \phi_k(x))}{(\phi_{k-1}(x), \phi_{k-1}(x))} \ .$$

Os polinômios $\phi_0(x), \phi_1(x), \phi_2(x), \ldots,$ *assim definidos, são dois a dois ortogonais, isto é, satisfazem (11.17).*

Prova: Faremos a prova por indução.

a) Inicialmente, temos que: $\phi_1(x)$ e $\phi_0(x)$ são ortogonais. De fato:

$$\begin{aligned}(\phi_1(x), \phi_0(x)) &= \left(x - \frac{(x,1)}{(1,1)}\, 1, 1\right) = \\ &= (x,1) - \frac{(x,1)}{(1,1)}\,(1,1) = 0\,.\end{aligned}$$

b) Supondo $(\phi_i(x), \phi_j(x)) = 0$, para $i \neq j,\ i,j = 0,1,\ldots,k$.

c) Provemos que $(\phi_{k+1}(x), \phi_j(x)) = 0 \quad , \quad j = 0,1,\ldots,k$.

Dividiremos a prova em três partes.

c.1) Consideremos inicialmente $j = k$. Temos então:

$$\begin{aligned}(\phi_{k+1}(x), \phi_k(x)) &= (x\phi_k(x) - \alpha_k\phi_k(x) - \beta_k\phi_{k-1}(x), \phi_k(x)) \\ &= (x\phi_k(x), \phi_k(x)) - \alpha_k\,(\phi_k(x), \phi_k(x)) - \beta_k\,(\phi_{k-1}(x), \phi_k(x)) \\ &= (x\phi_k(x), \phi_k(x)) - \frac{(x\phi_k(x), \phi_k(x))}{(\phi_k(x), \phi_k(x))}\,(\phi_k(x), \phi_k(x)) = 0.\end{aligned}$$

desde que, pela hipótese de indução, $\phi_{k-1}(x)$ e $\phi_k(x)$ são ortogonais, e onde substituímos α_k pelo valor dado no teorema.

c.2) Para $j = k-1$, temos:

$$\begin{aligned}(\phi_{k+1}(x), \phi_{k-1}(x)) &= (x\phi_k(x) - \alpha_k\phi_k(x) - \beta_k\phi_{k-1}(x), \phi_{k-1}(x)) \\ &= (x\phi_k(x), \phi_{k-1}(x)) - \alpha_k\,(\phi_k(x), \phi_{k-1}(x)) - \beta_k\,(\phi_{k-1}(x), \phi_{k-1}(x)) \\ &= (\phi_k(x), x\phi_{k-1}(x)) - \beta_k\,(\phi_{k-1}(x), \phi_{k-1}(x))\ ,\end{aligned}$$

desde que estamos utilizando o produto escalar definido por (11.18), e novamente aplicamos a hipótese de indução.

Mas,

$$(\phi_k(x), x\phi_{k-1}(x)) = (\phi_k(x), \phi_k(x))\ .$$

De fato, usando (11.19), temos, para $k = 1$, que:

$$\begin{aligned}(\phi_1(x), x\phi_0(x)) &= (\phi_1(x), x) \\ &= \left(\phi_1(x), \phi_1(x) + \frac{(x\phi_0(x), \phi_0(x))}{(\phi_0(x), \phi_0(x))}\ \phi_0(x)\right) \\ &= (\phi_1(x), \phi_1(x)) + \frac{(x\phi_0(x), \phi_0(x))}{(\phi_0(x), \phi_0(x))}\ (\phi_1(x), \phi_0(x)) \\ &= (\phi_1(x), \phi_1(x))\,,\end{aligned}$$

desde que $\phi_1(x)$ e $\phi_0(x)$ são ortogonais; e, para $k > 1$,

$$\begin{aligned}(\phi_k(x), x\phi_{k-1}(x)) &= (\phi_k(x), \phi_k(x) + \alpha_{k-1}\phi_{k-1}(x) + \beta_{k-1}\phi_{k-2}(x)) \\ &= (\phi_k(x), \phi_k(x)) \ .\end{aligned}$$

desde que pela hipótese de indução $\phi_{k-1}(x)$ e $\phi_{k-2}(x)$ são ortogonais. Então, finalmente, temos:

$$(\phi_{k+1}(x), \phi_{k-1}(x)) = (\phi_k(x), \phi_k(x)) - \beta_k\,(\phi_{k-1}(x), \phi_{k-1}(x)) = 0 \ ,$$

desde que

$$\beta_k = \frac{(\phi_k(x) - \phi_k(x))}{\phi_{k-1}(x), \phi_{k-1}(x))} \ .$$

c.3) Vamos considerar agora $j = k-2, k-3, \ldots, 1, 0$. Assim:

$$\begin{aligned}(\phi_{k+1}(x), \phi_j(x)) &= (x\phi_k(x) - \alpha_k\phi_k(x) - \beta_k\phi_{k-1}(x), \phi_j(x)) \\ &= (\phi_k(x), x\phi_j(x)) \ - \ \alpha_k\,(\phi_k(x), \phi_j(x)) \ - \ \beta_k\,(\phi_{k-1}(x), \phi_j(x)) \\ &= (\phi_k(x), x\phi_j(x)) \\ &= (\phi_k(x), \phi_{j+1}(x) + \alpha_j\phi_j(x) + \beta_j\phi_{j-1}(x)) \\ &= (\phi_k(x), \phi_{j+1}(x)) \ + \ \alpha_j\,(\phi_k(x), \phi_j(x)) \ + \ \beta_j\,(\phi_k(x), \phi_{j-1}(x)) \\ &= 0 \ , \ \ \text{pois} \ \ j < k-1.\end{aligned}$$

Portanto, os polinômios definidos em (11.19) são dois a dois ortogonais.

Observe que obter uma sequência de polinômios ortogonais pelo Teorema 11.4 é muito mais fácil do que obtê-los através do processo de ortogonalização de Gram-Schmidt. Muito mais fácil, pois o Teorema anterior fornece uma fórmula de recorrência que envolve apenas três termos para obter qualquer polinômio da sequência de grau $k \geq 2$, enquanto que o processo de Gram-Schmidt requer o cálculo de $k-1$ termos para obter o polinômio de grau k, $k = 1, 2, \ldots$.

11.3.1 Principais Polinômios Ortogonais

A sequência de polinômios $\phi_0(x), \phi_1(x), \phi_2(x), \ldots$, evidentemente, depende do produto escalar adotado (forma geral (11.18)). Os mais conhecidos (inclusive já tabelados) e com os quais trabalharemos são os seguintes:

Polinômios de Legendre

Os polinômios de Legendre $P_0(x), P_1(x), \ldots$, são obtidos segundo o produto escalar:

$$(f, g) = \int_{-1}^{1} f(x)\ g(x)\ d(x) \ , \qquad (11.20)$$

isto é, com $\omega(x) = 1$, $a = -1$ e $b = 1$.

Polinômios de Tchebyshev

O produto escalar usado para obter os polinômios de Tchebyshev $T_0(x), T_1(x), \ldots$ é dado por:

$$(f, g) \; = \; \int_{-1}^{1} \frac{1}{\sqrt{1 - x^2}} \; f(x) \; g(x) \; dx \; , \tag{11.21}$$

ou seja, $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}, \; a = -1$ e $b = 1$.

Polinômios de Laguerre

Os polinômios de Laguerre $L_0(x), L_1(x), \ldots$, são obtidos usando-se o produto escalar:

$$(f, g) \; = \; \int_{0}^{\infty} e^{-x} \; f(x) \; g(x) \; dx, \tag{11.22}$$

portanto, $\omega(x) = e^{-x}, \; a = 0$ e $b = \infty$.

Polinômios de Hermite

Obtemos os polinômios de Hermite, usando o produto escalar:

$$(f, g) \; = \; \int_{-\infty}^{\infty} e^{-x^2} \; f(x) \; g(x) \; dx \; , \tag{11.23}$$

isto é, $\omega(x) = e^{-x^2}, \; a = -\infty$ e $b = \infty$.

Exemplo 11.8 - *Obter os primeiros polinômios de Legendre.*

Solução: Para obter os polinômios de Legendre, devemos utilizar o Teorema 11.4 e o produto escalar definido por (11.20). Assim:

$$
\begin{aligned}
P_0(x) &= 1 \\
P_1(x) &= x - \frac{(x,1)}{(1,1)} \cdot 1 \\
&= x - \frac{\int_{-1}^{1} x \; dx}{\int_{-1}^{1} \; dx} \\
&= x - \left. \frac{x^2/2}{x} \right]_{-1}^{1} \\
\Rightarrow P_1(x) &= x \; . \\
P_2(x) &= x \; P_1(x) - \alpha_1 P_1(x) \; - \; \beta_1 P_0(x), \quad \text{onde} \\
\alpha_1 &= \frac{(xP_1(x), P_1(x))}{(P_1(x), P_1(x))} = \frac{\int_{-1}^{1} \; x^3 \; dx}{\int_{-1}^{1} \; x^2 \; dx} \\
&= \left. \frac{x^4/4}{x^3/3} \right]_{-1}^{1} = 0 \; , \\
\beta_1 &= \frac{(P_1(x), P_1(x))}{(P_0(x), P_0(x))} = \frac{\int_{-1}^{1} \; x^2 \; dx}{\int_{-1}^{1} \; dx} \\
&= \left. \frac{x^3/3}{x} \right]_{-1}^{1} = \frac{2/3}{2} = \frac{1}{3} \; , \\
\Rightarrow P_2(x) &= x^2 + 0 \times x - \frac{1}{3} \times 1 = x^2 \; - \; \frac{1}{3}. \\
&\vdots
\end{aligned}
$$

11.3.2 Propriedades dos Polinômios Ortogonais

Vejamos algumas das propriedades dos polinômios ortogonais que serão importantes para a obtenção das fórmulas de Quadratura de Gauss.

Propriedade 1 - Sejam $\phi_0(x), \phi_1(x), \phi_2(x), \ldots$, polinômios ortogonais, não nulos, segundo um produto escalar qualquer. Então, qualquer polinômio de grau menor ou igual a n pode ser escrito como combinação linear de $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$.

Prova: Os polinômios $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$ constituem uma base para o espaço dos polinômios de

grau menor ou igual a n. Assim, se $Q(x)$ é um polinômio da forma:

$$Q(x) = a_0 + a_1 x + \ldots + a_n x^n ,$$

então $Q(x)$ pode ser escrito, através de mudança de base, como:

$$Q(x) = b_0\phi_0(x) + b_1\phi_1(x) + \ldots + b_n\phi_n(x) .$$

Propriedade 2 - Sejam $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$ nas condições da propriedade 1. Então $\phi_n(x)$ é ortogonal a qualquer polinômio $Q(x)$ de grau menor que n.

Prova: Seja $Q(x)$, um polinômio de grau $n-1$. Pela propriedade anterior temos que:

$$Q(x) = b_0\phi_0(x) + b_1\phi_1(x) + \ldots + b_{n-1}\phi_{n-1}(x) ,$$

então:

$$\begin{aligned}(Q(x), \phi_n(x)) &= (b_0\phi_0(x) + b_1\phi_1(x) + \ldots + b_{n-1}\phi_{n-1}(x), \phi_n(x)) \\ &= b_0(\phi_0(x), \phi_n(x)) + b_1(\phi_1(x), \phi_n(x)) + \ldots + b_{n-1}(\phi_{n-1}(x), \phi_n(x)) \\ &= 0.\end{aligned}$$

desde que os polinômios $\phi_0(x), \phi_1(x), \ldots, \phi_n(x)$ são dois a dois ortogonais.

Propriedade 3 - Sejam $\phi_0(x), \phi_1(x), \phi_2(x), \ldots$, polinômios ortogonais segundo o produto escalar:

$$(f, g) = \int_a^b \omega(x)\ f(x)\ g(x)\ dx,$$

com $\omega(x) \geq 0$ e contínua em $[a, b]$. Então $\phi_n(x)$ possui n raízes (reais) distintas em $[a, b]$.

Prova: Para verificar a veracidade desta propriedade dividiremos a prova em três partes, isto é, provaremos que:

a) $\phi_n(x)$ possui algum zero em $[a, b]$,

b) os zeros de $\phi_n(x)$ em $[a, b]$, são simples,

c) os n zeros de $\phi_n(x)$ estão em $[a, b]$.

Os três itens serão provados por absurdo. Assim, para provar **a)**, vamos supor, por absurdo, que $\phi_n(x)$ não possui zeros em $[a, b]$. Portanto em $[a, b]$, $\phi_n(x) \neq 0$. Assim:

$$\begin{aligned}(\phi_n(x), \phi_0(x)) &= \int_a^b \omega(x)\phi_n(x)\ \phi_0(x)dx \\ &= \int_a^b \omega(x)\phi_n(x)\ dx \neq 0 ,\end{aligned}$$

desde que $\phi_0(x) = 1,\ \omega(x) \geq 0$, mas não pode ser identicamente nula, e $\phi_n(x) \neq 0$ em $[a, b]$.

Mas $\phi_n(x)$ e $\phi_0(x)$ são ortogonais, e portanto $(\phi_n(x), \phi_0(x)) = 0$. Logo é um absurdo supor que $\phi_n(x)$ não possui zeros em $[a, b]$.

Para provar **b)** vamos supor, por absurdo, que exista uma raiz de $\phi_n(x)$ que seja de multiplicidade 2. Seja x_1 essa raiz. Portanto:

$$\frac{\phi_n(x)}{(x - x_1)^2} \; ,$$

é um polinômio de grau $n - 2$. Assim, pela propriedade 2:

$$(\phi_n(x), \frac{\phi_n(x)}{(x - x_1)^2}) \; = \; 0.$$

Mas, usando as propriedades de produto escalar, obtemos:

$$\begin{aligned} (\phi_n(x), \frac{\phi_n(x)}{(x - x_1)^2}) \; &= \; \int_a^b \; \omega(x)\phi_n(x) \; \frac{\phi_n(x)}{(x - x_1)^2} \; dx \\ &= \; \int_a^b \; \omega(x) \frac{\phi_n(x)}{(x - x_1)} \; \frac{\phi_n(x)}{(x - x_1)} \; dx \\ &= \; (\frac{\phi_n(x)}{(x - x_1)}, \frac{\phi_n(x)}{(x - x_1)}) \; \geq \; 0, \end{aligned}$$

onde a igualdade é válida se e somente se $\frac{\phi_n(x)}{(x - x_1)}$ for o polinômio nulo. Portanto é um absurdo supor que os zeros de $\phi_n(x)$ em $[a, b]$ não são simples.

Finalmente, para provar **c)** vamos supor, por absurdo, que exista apenas j zeros de $\phi_n(x)$ em $[a, b]$, com $j < n$. Sejam $x_1, x_2, \ldots, x_j$ os zeros de $\phi_n(x)$ em $[a, b]$. Então, podemos escrever:

$$\phi_n(x) \; = \; (x - x_1)(x - x_2) \ldots (x - x_j) q_{n-j}(x)$$

onde $q_{n-j} \neq 0$ em $[a, b]$. Assim, pela propriedade 2, segue que:

$$(\phi_n(x), (x - x_1)(x - x_2) \ldots (x - x_j)) \; = \; 0.$$

Mas, usando as propriedades de produto escalar, temos que:

$$(\phi_n(x), (x - x_1)(x - x_2) \ldots (x - x_j)) \; =$$

$$\begin{aligned} &= \; \int_a^b \; \omega(x)(x - x_1) \ldots (x - x_j) q_{n-j}(x)(x - x_1) \ldots (x - x_j) \; dx \\ &= \; \int_a^b \; \omega(x)(x - x_1)^2 \ldots (x - x_j)^2 \; q_{n-j}(x) \; dx \; \neq \; 0. \end{aligned}$$

Portanto, é um absurdo supor que os n zeros de $\phi_n(x)$ não estão em $[a, b]$.

Assim, acabamos de provar que $\phi_n(x)$ possui n zeros (reais), distintos em $[a, b]$.

Propriedade 4 - Sejam $\phi_0(n), \phi_1(x), \phi_2(x), \ldots$, nas condições da propriedade 3. Sejam $x_0, x_1, \ldots, x_n$ as raízes de $\phi_{n+1}(x)$. Se $f(x)$ é um polinômio de grau menor ou igual a $2n+1$, então:

$$\int_a^b \omega(x)\ f(x)\ dx \ = \ \sum_{k=0}^{n} A_k\ f(x_k) \ , \tag{11.24}$$

onde

$$A_k \ = \ \int_a^b \omega(x)\ \ell_k(x)\ dx.$$

Prova: Como $x_0, x_1, \ldots, x_n$ são raízes de $\phi_{n+1}(x)$, podemos escrever:

$$\phi_{n+1}(x) \ = \ a_0\,(x-x_0)\,(x-x_1)\ldots(x-x_n)\,. \tag{11.25}$$

Seja $P_n(x)$ o polinômio de interpolação de $f(x)$ sobre $x_0, x_1, \ldots, x_n$ em $[a,b]$. Sabemos que:

$$f(x) \ = \ P_n(x) \ + \ R_n(x) \ ,$$

onde $R_n(x)$ é o erro na interpolação. Assim:

$$f(x) - P_n(x) = R_n(x) \ = \ (x-x_0)\,(x-x_1)\ldots(x-x_n)\ \frac{f(n+1)(\xi)}{(n+1)!} \ ,$$

com $a \leq \xi \leq b$ e ξ dependendo de x.

Então, em vista de (11.25) e de que ξ é função de x, podemos escrever:

$$f(x) - P_n(x) \ = \ b_0\ \phi_{n+1}(x)\ \frac{f^{(n+1)}(x)}{(n+1)!} \ .$$

Como $f(x)$ é um polinômio de grau menor ou igual a $2n+1$, temos que:

$$q(x) \ = \ \frac{f^{(n+1)}(x)}{(n+1)!} \ ,$$

é um polinômio de grau menor ou igual a n. Assim, podemos escrever:

$$f(x) \ - \ P_n(x) \ = \ b_0\ \phi_{n+1}(x)\ q(x) \ . \tag{11.26}$$

Integrando (11.26) de a até b, com a função peso $\omega(x)$, obtemos que:

$$\int_a^b \omega(x)\ [f(x) \ - \ P_n(x)]\ dx \ = \ \int_a^b \omega(x)\ b_0\ \phi_{n+1}(x)\ q(x)\ dx \ .$$

Pela propriedade 2, o lado direito da igualdade acima é igual a zero. Assim:

$$\int_a^b \omega(x)\ [f(x) \ - \ P_n(x)]\ dx \ = \ 0 \ ,$$

ou

$$\begin{aligned}
\int_a^b \omega(x)\ f(x)\ dx &= \int_a^b \omega(x)\ P_n(x)\ dx \\
&= \int_a^b \omega(x) \left[\sum_{k=0}^{n} \ell_k(x)\ f(x_k) \right] dx \\
&= \sum_{k=0}^{n} f(x_k) \int_a^b \omega(x)\ \ell_k(x)\ dx \\
&= \sum_{k=0}^{n} A_k f(x_k)\ .
\end{aligned}$$

Portanto, fica provada a relação (11.24).

Esta propriedade garante então que, para integrar um polinômio de um certo grau k, basta trabalharmos com um polinômio ortogonal de grau aproximadamente k/2. E mais, descartados os erros de arredondamento, o resultado deve ser exato.

11.4 Fórmulas de Quadratura de Gauss

São fórmulas usadas para se calcular:

$$\int_a^b \omega(x)\ f(x)\ dx\ ,$$

valendo-se do resultado da propriedade 4. Calculamos o valor aproximado da integral usando:

$$\int_a^b \omega(x)\ f(x)\ dx \simeq \sum_{k=0}^{n} A_k\ f(x_k)\ ,$$

onde:

$$A_k = \int_a^b \omega(x)\ \ell_k(x)\ dx\ ,$$

e $\ell_k(x)$ são os polinômios de Lagrange sobre as raízes $x_0, x_1, \ldots, x_n$ de $\phi_{n+1}(x)$.

Assim, o **procedimento** para se calcular uma integral usando **Quadratura de Gauss**, é o seguinte:

a) determinar o polinômio ortogonal $\phi_{n+1}(x)$, segundo o produto escalar conveniente, isto é, com a função peso $\omega(x)$ e no intervalo $[a, b]$.

b) calcular as raízes $x_0, x_1, \ldots, x_n$ de $\phi_{n+1}(x)$.

c) determinar os polinômios de Lagrange $\ell_k(x)$, $k = 0, 1, \ldots, n$, usando os pontos $x_0, x_1, \ldots, x_n$ obtidos em b).

d) calcular $A_k = \int_a^b \omega(x)\ \ell_k(x)\ dx$, $k = 0, 1, \ldots, n$.

e) calcular o valor de $f(x)$ em $x_0, x_1, \ldots, x_n$.

f) calcular, finalmente,

$$\int_a^b \omega(x)\ f(x)\ dx \ \simeq \ \sum_{k=0}^{n} A_k\ f\left(x_k\right)\ .$$

Exemplo 11.9 - *Usando quadratura de Gauss, calcular:*

$$\int_{-1}^{1} (x^3 - 5x)\ dx\ .$$

Solução: É claro que para resolver esta integral não precisamos de método numérico, entretanto este exemplo servirá para ilustrar como resolver uma integral usando o procedimento dado anteriormente.

Na integral, vemos que $a = -1,\ b = 1,\ \omega(x) = 1,$ e $f(x) = x^3 - 5x$. Assim $f(x)$ é um polinômio de grau 3, e pela propriedade 4, temos que: se $f(x)$ é um polinômio de grau $2n+1$, o resultado da integral é exato (a menos de erros de arredondamento).

Portanto, fazendo $2n + 1 = 3$, obtemos que $n = 1$.

Assim, devemos utilizar os zeros de $\phi_{n+1}(x) = \phi_2(x)$, para resolver a integral. O produto escalar, para obter $\phi_2(x)$, será: $\int_{-1}^{1}\ f(x)\ g(x)\ dx$. Logo o polinômio procurado é $x^2 - \frac{1}{3}$, (ver exemplo 11.8).

Portanto, fazendo $x^2 - \frac{1}{3}\ =\ 0$, obtemos $x_0 = -0.57735$ e $x_1 = 0.55735$, (que são os zeros de $\phi_2(x)$ em $[-1, 1]$).

Temos que:

$$\ell_0(x) \ = \ \frac{(x - x_1)}{(x_0 - x_1)}, \qquad \ell_1(x) \ = \ \frac{(x - x_0)}{(x_1 - x_0)}\ ,$$

e portanto:

$$\begin{aligned}
A_0 \ &= \ \int_{-1}^{1} \ell_0(x)\ dx = \int_{-1}^{1} \frac{(x - x_1)}{(x_0 - x_1)}\ dx \\
&= \ \frac{1}{(x_0 - x_1)} \int_{-1}^{1} (x - x_1)\ dx = \frac{1}{(x_0 - x_1)} (\frac{x^2}{2} - (x\ x_1))\ ]_{-1}^{1} \\
&= \ \frac{-2x_1}{-2x_1}\ =\ 1\ ,
\end{aligned}$$

desde que $x_0 = -x_1$ e $\frac{x^2}{2}\ ]_{-1}^{1}\ =\ 0$. Do mesmo modo:

$$\begin{aligned}
A_1 \ &= \ \int_{-1}^{1} \ell_1(x)\ dx = \int_{-1}^{1} \frac{(x - x_0)}{(x_1 - x_0)}\ dx \\
&= \ \frac{1}{(x_1 - x_0)} \int_{-1}^{1} (x - x_0)\ dx = \frac{1}{(x_1 - x_0)} (\frac{x^2}{2} - (x\ x_0))\ ]_{-1}^{1} \\
&= \ \frac{-2x_0}{-2x_0}\ =\ 1\ .
\end{aligned}$$

Agora, calculamos a f nos zeros de $\phi_2(x)$. Assim:

$$\begin{aligned} f(x_0) &= f(-0.57735) = (-0.57735)^3 - 5(-0.57735)\ , \\ f(x_1) &= f(0.57735) = (0.57735)^3 - 5(0.57735). \end{aligned}$$

Finalmente, podemos calcular a integral, isto é:

$$\begin{aligned} \int_{-1}^{1} x^3 - 5x\ dx &= A_0\ f(x_0) + A_1\ f(x_1) \\ &= 1 \times [(-0.57735)^3 - 5(-0.57735)] \\ &+ 1 \times [(0.57735)^3 - 5(0.57735)] = 0\ . \end{aligned}$$

Observação: O procedimento dado é válido para qualquer produto escalar. Quando particularizamos o produto escalar aos já mencionados anteriormente, isto é, quando usamos os produtos escalares para obter os polinômios de Legendre, Tchebyshev, Laguerre e Hermite, necessitamos apenas efetuar os passos **e)** e **f)** do procedimento dado anteriormente, pois os valores de x_k e A_k já estão tabelados, (as tabelas encontram-se no final deste capítulo). Neste caso, temos as **fórmulas de quadratura de Gauss-Legendre, Gauss-Tchebyshev, Gauss-Laguerre e Gauss-Hermite**.

Antes de descrevermos cada uma destas fórmulas daremos as

INSTRUÇÕES PARA O USO DAS TABELAS

1 - Tabelas 1, 2 e 4

a) Os valores de x_i e A_i são apresentados na forma normalizada, isto é, na forma $0.\ \ldots \times 10^j$, onde j aparece, entre parêntesis, antes do número. Quando não aparecer a potência, significa $j = 0$.

b) Cada valor é dado com 10 casas decimais. No livro de [Stroud, 19..], esses valores são apresentados com 30 casas decimais, agrupadas de 10 em 10 para facilitar a leitrura.

c) Nas tabelas **N** significa o número de pontos que devemos tomar para resolver a integral. Observe que da teoria tiramos o valor de **n** o que implica que devemos tomar $\mathbf{n+1}$ pontos. Assim $\mathbf{N = n + 1}$.

d) Como o intervalo de integração é simétrico em relação à origem, as raízes x_i também o são. Na tabela aparece apenas x_i sem sinal; então, devemos considerar $\pm\ x_i$. Por exemplo: tabela 1 : $N = 3$, temos $n = 2$, isto é, x_0, x_1 e x_2. Neste caso $x_0 = -0.77459\ldots$; $x_1 = 0$ e $x_2 = +0.77459\ldots$.

e) Os A_i são sempre positivos! Observe que devemos tomar para A_i os valores que estão na mesma linha dos x_i, isto é, considerando os pontos x_i do item **d)** teremos: $A_0 = 0.55555\ldots$, $A_1 = 0.88888\ldots$, e $A_2 = 0.55555\ldots$.

2 - Tabela 3

Uso análogo às tabelas 1, 2 e 4 com uma única exceção:

Os x_i são sempre positivos e estão todos tabelados. Por exemplo, para $N = 3$ aparecem 3 valores de x_i.

Observe que em todas as tabelas podemos tomar os valores de x_i em qualquer ordem, mas para cada x_i devemos tomar o valor A_i correspondente.

Daremos a seguir as fórmulas de quadratura de Gauss-Legendre, Gauss-Tchebyshev, Gauss-Laguerre e Gauss-Hermite. Assim:

11.4.1 Fórmula de Gauss-Legendre

Para utilizar a fórmula de Gauss-Legendre a integral a ser calculada deve ter a função peso $\omega(x) = 1$, $a = -1$ e $b = 1$. Caso o intervalo de integração não coincida com o intervalo $[-1, 1]$, devemos fazer uma mudança de variável. Daremos a seguir alguns exemplos.

Exemplo 11.10 - *Resolver a integral do exemplo anterior, isto é, calcular:*

$$\int_{-1}^{1} (x^3 - 5x)\ dx\ ,$$

usando quadratura de Gauss.

Solução: Como $\omega(x) = 1$, $a = -1$ e $b = 1$, estamos nas condições de utilizar a fórmula de Gauss-Legendre. Pelo exemplo 11.9, vemos que devemos utilizar os zeros de $\phi_2(x)$, desde que $f(x)$ é um polinômio de grau 3. Assim, pela Tabela 1, (ao final deste capítulo), com $\mathbf{N} = \mathbf{n} + \mathbf{1} = 2$, temos que:

$$\begin{aligned} x_0 &= -0.55735\ , \quad A_0 = (1)0.10000 = 1 = A_1\ , \\ x_1 &= 0.55735\ . \end{aligned}$$

Portanto:

$$\begin{aligned} f(x_0) &= f(-0.57735) = (-0.57735)^3 - 5(-0.57735)\ , \\ f(x_1) &= f(0.57735) = (0.57735)^3 - 5(0.57735). \end{aligned}$$

Finalmente, podemos calcular a integral. Assim:

$$\begin{aligned} \int_{-1}^{1} x^3 - 5x\ dx &= A_0\ f(x_0) + A_1\ f(x_1) \\ &= 1 \times [(-0.57735)^3 - 5(-0.57735)] \\ &+ 1 \times [(0.57735)^3 - 5(0.57735)] = 0\ . \end{aligned}$$

Exemplo 11.11 - *Calcular:*

$$\int_1^2 \frac{dx}{x} \; ,$$

usando quadratura de Gauss-Legendre.

Solução: Como o intervalo de integração não coincide com o intervalo [-1,1] de definição do produto escalar de Legendre, devemos fazer uma mudança de variável. Assim:

$$\begin{array}{lllll} & \text{quando} & x = 1 & \text{devemos ter} & t = -1 \; , \\ \text{e} & & & & \\ & \text{quando} & x = 2 & \text{devemos ter} & t = 1 \; . \end{array}$$

A equação da reta que passa por $(1,2)$ e $(-1,1)$ pode ser obtida calculando-se o seguinte determinante e igualando-o a zero, isto é:

$$\begin{vmatrix} x & t & 1 \\ 1 & -1 & 1 \\ 2 & 1 & 1 \end{vmatrix} = -2x + t + 3 = 0 \; .$$

Resolvendo esta equação, obtemos:

$$x = \frac{t+3}{2} \; , \quad \text{e} \quad dx = \frac{1}{2} \, dt.$$

Portanto:

$$\int_1^2 \frac{dx}{x} = \int_{-1}^1 \frac{1}{\left(\frac{t+3}{2}\right)} \frac{1}{2} \, dt = \int_{-1}^1 \frac{dt}{t+3} \; .$$

Estão, agora, satisfeitas as condições da fórmula de quadratura de Gauss-Legendre com $\omega(x) = 1, \; a = -1, \; b = 1$ e $f(t) = \frac{1}{t+3}$.

Como $f(t)$ não é um polinômio, não temos maiores informações sobre o grau do polinômio a ser usado. Assim, resolveremos esta integral fixando $\mathbf{n} = \mathbf{2}$. Mais adiante nos preocuparemos com a precisão do resultado.

Da Tabela 1, com $\mathbf{N} = \mathbf{n} + \mathbf{1} = 3$, obtemos:

$$\begin{array}{ll} t_0 = -0.7746 \; , & A_0 = 0.5556 = A_2 \; , \\ t_1 = 0 \; , & A_1 = 0.8889 \; , \\ t_2 = 0.7746 \; . & \end{array}$$

Assim:

$$f(t_0) = \frac{1}{-0.7746 + 3} = 0.4494 \; ,$$

$$f(t_1) = \frac{1}{0 + 3} = 0.3333 \; ,$$

$$f(t_2) = \frac{1}{0.7746 + 3} = 0.2649 \; .$$

Logo:

$$\begin{aligned}\int_{-1}^{1} \frac{1}{t+3}\, dt &\simeq \sum_{k=0}^{2} A_k\ f(t_k)\\ &= (0.4494 + 0.2649)(0.5556) + (0.3333)(0.8889)\\ &= 0.6931\ .\end{aligned}$$

Portanto:

$$\int_{1}^{2} \frac{dx}{x} \simeq 0.6931.$$

Observe que:

a) A solução exata é:

$$\int_{1}^{2} \frac{dx}{x} = \ln\ 2 = 0.693147180\ldots$$

b) Usando quadratura de Gauss, com $\mathbf{n} = 4$ e 9 casas decimais obtemos:

$$\int_{1}^{2} \frac{dx}{x} \simeq 0.693147156\ .$$

que pode ser considerado um bom resultado.

11.4.2 Fórmula de Gauss-Tchebyshev

Para utilizar as fórmulas de Gauss-Tchebyshev a integral a ser calculada deve ter a função peso $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}$, $a = -1$ e $b = 1$. Novamente, caso o intervalo de integração não coincida com o intervalo $[-1, 1]$, devemos fazer uma mudança de variável. Daremos a seguir alguns exemplos.

Exemplo 11.12 - *Usando quadratura de Gauss, calcular:*

$$\int_{-1}^{1} \frac{sen\ x}{\sqrt{1-x^2}}\ dx\ .$$

Solução: Temos que: $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}$, $a = -1$, $b = 1$ e $f(x) = sen\ x$. Portanto estamos nas condições de utilizar a fórmula de Gauss-Tchebyshev. Como $f(x)$ não é um polinômio não temos como obter o número exato de pontos para calcular a integral. Assim, fixemos $\mathbf{n} = 2$. Da tabela 2, com $\mathbf{N} = \mathbf{n} + \mathbf{1} = 3$, obtemos:

$$\begin{aligned} x_0 &= -0.8860\ ,\\ x_1 &= 0\ , \qquad A_0 = A_1 = A_2 = (1)0.10472 = 1.0472\ ,\\ x_2 &= 0.8860\ .\end{aligned}$$

Coloque sua máquina de calcular em radianos para calcular o valor da f nos pontos x_i, i = 0, 1, 2. Assim:

$$f(x_0) = f(-0.8860) = sen\,(-0.8860) = -0.7617 = -f(x_2)\ ,$$

$$f(x_1) = f(0) = sen\ 0 = 0\ .$$

Portanto:

$$\int_{-1}^{1} \frac{sen\ x}{\sqrt{1-x^2}}\ dx = 1.0472\ (-0.7617 + 0 + 0.7617) = 0.$$

Exemplo 11.13 - *Calcular:*

$$\int_{-2}^{2} \frac{t^3 + 2t^2}{4\ \sqrt{4-t^2}}\ dt\ ,$$

exatamente, a menos de erros de arredondamento, usando quadratura de Gauss.

Solução: Como o intervalo de integração não coincide com o intervalo $[-1, 1]$, devemos fazer uma mudança de variável. Assim:

$$\begin{array}{ll} & \text{quando} \quad t = -2 \quad \text{devemos ter} \quad x = -1\ , \\ \text{e} & \\ & \text{quando} \quad t = 2 \quad \text{devemos ter} \quad x = 1\ . \end{array}$$

Novamente, a equação da reta que passa por $(-2, 2)$ e $(-1, 1)$ pode ser obtida calculando-se o seguinte determinante e igualando-o a zero, isto é:

$$\begin{vmatrix} t & x & 1 \\ -2 & -1 & 1 \\ 2 & 1 & 1 \end{vmatrix} = t - 2x = 0\ .$$

Resolvendo esta equação, obtemos:

$$t = 2x \quad \text{e} \quad dt = 2\ dx.$$

Portanto:

$$\begin{aligned} \int_{-2}^{2} \frac{t^3 + 2t^2}{4\ \sqrt{4-t^2}}\ dt &= \int_{-1}^{1} \frac{(2x)^3 + 2(2x)^2}{4\ \sqrt{4-(2x)^2}}\ 2\ dx \\ &= 2 \int_{-1}^{1} \frac{8x^3 + 8x^2}{8\ \sqrt{1-x^2}}\ dx \\ &= 2 \int_{-1}^{1} \frac{x^3 + x^2}{\sqrt{1-x^2}}\ dx\ . \end{aligned}$$

Assim temos: $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}$, $a = -1, b = 1$ e $f(x) = x^3 + x^2$, e portanto estamos nas condições da fórmula de Gauss-tchebyshev.

Vemos que f é um polinômio, o que era de se esperar, pois queremos calcular a integral exatamente, (a menos de erros de arredondamento). Fazendo $2n + 1 = 3$ obtemos que $\mathbf{n} = 1$, ou seja, devemos tomar 2 pontos. Da tabela 2, com $\mathbf{N} = 2$, obtemos:

$$x_0 = -0.7071\ , \qquad A_0 = A_1 = (1)0.15708 = 1.5708\ ,$$

$$x_1 = 0.7071\ .$$

Agora:

$$f(x_0) = (-0.7071)^3 + (-0.7071)^2 = 0.1464 ,$$

$$f(x_1) = (0.7071)^3 + (0.7071)^2 = 0.8535 .$$

Portanto:

$$\int_{-2}^{2} \frac{t^3 + 2t^2}{4\ \left(4 - t^2\right)^{1/2}}\ dt = 2 \int_{-1}^{1} \frac{x^3 + x^2}{\left(1 - x^2\right)^{1/2}}\ dx$$

$$= 2 \times 1.5708\ (0.1464 + 0.8535)$$

$$= 3.1413.$$

Observe que sempre falamos em obter o resultado exato, (a menos de erros de arredondamento), pois não trabalhamos com todas as casas decimais disponíveis nos x_i e nos A_i. Assim, teoricamente, o resultado deveria ser exato, mas na prática ele depende da quantidade de dígitos significativos com que trabalhamos.

11.4.3 Fórmula de Gauss-Laguerre

Para utilizar a fórmula de Gauss-Laguerre a integral a ser calculada deve ter a função peso $\omega(x) = e^{-x}$, $a = 0$ e $b = \infty$. Novamente, caso o intervalo de integração não coincida com o intervalo $[0, \infty]$, devemos fazer uma mudança de variável. Daremos a seguir alguns exemplos.

Exemplo 11.14 - *Usando quadratura de Gauss, calcular:*

$$\int_{0}^{\infty} e^{-x}\ \cos\ x\ dx.$$

Solução: Como $\omega(x) = e^{-x}$, $a = 0$ e $b = \infty$, estamos nas condições da fórmula de quadratura de Gauss-Laguerre. Desde que $f(x) = \cos\ x$, não temos condições de determinar o número exato de pontos que devemos tomar para resolver a integral. Assim fixemos $\mathbf{n} = 2$. Portanto pela Tabela 3, com $\mathbf{N} = 3$, obtemos:

$$x_0 = 0.4158 , \qquad A_0 = 0.7111 ,$$

$$x_1 = (1)0.22943 = 2.2943 , \qquad A_1 = 0.2785 ,$$

$$x_2 = (1)0.62900 = 6.2900 , \qquad A_2 = (-1)0.1039 = 0.01039 .$$

Agora:

$$f(x_0) = f(0.4158) = \cos(0.4158) = 0.9148 ,$$

$$f(x_1) = f(2.2943) = \cos(2.2943) = -0.6620 ,$$

$$f(x_2) = f(6.2900) = \cos(6.2900) = 1.0 .$$

Portanto:

$$\int_{0}^{\infty} e^{-x}\ \cos\ x\ dx = 0.7111\ (0.9148) + 0.2785\ (-0.6620) + 0.01039\ (1.0)$$

$$= 0.4765 .$$

Exemplo 11.15 - *Calcular, usando quadratura de Gauss,*

$$\int_1^\infty e^{-x} x^2 \ dx \ .$$

Solução: Aqui, o intervalo de integração e do produto escalar não coincidem. Então devemos fazer uma mudança de variável. Como um dos limites de integração é infinito não podemos utilizar a equação da reta para obter a mudança de variável. Nesse caso devemos fazer a mudança de variável via cálculo. Assim, se tomarmos $x = z + 1$, teremos:

$$\text{quando } x = 1 \ \rightarrow \ z = 0 \ ; \ \text{ e quando } x = \infty \ \rightarrow \ z = \infty \ ; \ dx = dz \ .$$

Logo:

$$\int_1^\infty e^{-x} x^2 \ dx \ = \ \int_0^\infty e^{-(z+1)} (z+1)^2 dz \ = \ e^{-1} \int_0^\infty e^{-z} (z+1)^2 \ dz \ ,$$

portanto nas condições da fórmula de quadratura de Gauss-Laguerre.

Como $f(z)$ é um polinômio de grau 2, fazendo $2n + 1 = 2$, obtemos que $n = \frac{1}{2} \rightarrow n = 1$, desde que n indica o índice do último ponto a ser considerado e portanto deve ser um inteiro. Assim da Tabela 3, com $\mathbf{N} = 2$, temos que:

$$z_0 = 0.5857864 \ , \qquad A_0 = 0.8535534 \ ,$$

$$z_1 = 3.4142136 \ , \qquad A_1 = 0.1464466 \ .$$

Calculando $f(z) = (z+1)^2$ para z_0 e z_1, obtemos:

$$f(z_0) = 2.5147185 \ ,$$

$$f(z_1) = 19.485282 \ .$$

Logo:

$$\sum_{k=0}^{1} A_k f(x_k) \ = \ A_0 f(z_0) \ + \ A_1 f(z_1) \ = \ 4.9999998.$$

Portanto:

$$\int_1^\infty e^{-x} x^2 dx = e^{-1} \int_0^\infty e^{-z} (z+1)^2 dz \ = \ e^{-1}(4.9999998) \ .$$

Observe que o resultado exato da integral é $\frac{5}{e}$. A pequena diferença que existe entre o resultado exato e o valor obtido é devida aos erros de arredondamento.

11.4.4 Fórmula de Gauss-Hermite

Para utilizar a fórmula de Gauss-Hermite a integral a ser calculada deve ter a função peso $\omega(x) = e^{-x^2}$; $a = -\infty$ e $b = \infty$. Neste caso se o intervalo de integração não coincidir com o intervalo $[-\infty, \infty]$, não podemos utilizar a fórmula de Gauss-Hermite. Daremos a seguir exemplo.

Exemplo 11.16 - *Calcular, usando quadratura de Gauss,*

$$\int_{-\infty}^\infty \frac{e^{-x^2}}{2} \ x^2 \ dx \ .$$

Solução: Estamos nas condições da fórmula de Gauss-Hermite com $f(x) = \frac{x^2}{2}$. Assim, pela propriedade 4, devemos tomar 2 pontos. Pela Tabela 4, com $\mathbf{N} = 2$, temos que:

$$x_0 = -0.7071 , \quad A_0 = A_1 = 0.8862 ,$$

$$x_1 = 0.7071 .$$

Assim:

$$f(x_0) = \frac{(-0.7071)^2}{2} = 0.25 = f(x_1) .$$

Portanto:

$$\int_{-\infty}^{\infty} \frac{e^{-x^2}}{2} x^2 \, dx = 0.8662 \times 2 \times 0.25 = 0.4431.$$

Observe que neste exemplo consideramos $f(x) = \frac{x^2}{2}$, mas poderíamos ter colocado $\frac{1}{2}$ fora da integral e considerado $f(x) = x^2$.

11.5 Erro nas Fórmulas de Gauss

Quando $f(x)$ é um polinômio, sabemos que as fórmulas de quadratura fornecem um resultado exato a menos, é claro, dos erros de arredondamento.

Na maioria das situações reais, $f(x)$ não é um polinômio e, portanto, sua integral é aproximada quando calculada através das fórmulas de quadratura.

Exibiremos algumas expressões do termo do resto (ou erro de truncamento) para as várias fórmulas apresentadas. Não nos preocuparemos com a dedução de tais expressões por ser extremamente trabalhosa e sem nenhum interesse prático.

Fórmula de Gauss-Legendre

A expressão do erro, para a fórmula de Gauss-Legendre, é dada por:

$$E_n = \frac{2^{2n+3}[(n+1)!]^4}{(2n+3)[(2n+2)!]^3} f^{(2n+2)}(\xi) \quad , \quad \xi \in (a,b) .$$

Fórmula de Gauss-Tchebyshev

A expressão do erro para a fórmula de Gauss-Tchebyshev é dada por:

$$E_n = \frac{2\pi}{e^{2n+2}(2n+2)!} f^{(2n+2)} (\xi) \quad , \quad \xi \in (a,b) .$$

Formula de Gauss-Laguerre

A expressão do erro, para a fórmula de Gauss-Laguerre, é dada por:

$$E_n = \frac{[(n+1)!]^2}{(2n+2)!} f^{(2n+2)} (\xi) \quad , \quad \xi \in (a,b) .$$

Fórmula de Gauss-Hermite

A expressão do erro, para a fórmula de Gauss-Hermite, é dada por:

$$E_n \; = \; \frac{(n+1)!\sqrt{\pi}}{2^{n+1}(2n+2)!} \; f^{(2n+2)} \; (\xi) \quad , \quad \xi \; \varepsilon \; (a,b) \; .$$

Como pode ser observado, todas as fórmulas do erro contém a derivada da f de ordem $2n+2$, onde n é o índice do último ponto considerado no cálculo da integral. Assim, usar a fórmula do erro para obter o número de pontos necessários para calcular a integral com uma determinada precisão torna-se inviável. Portanto, na prática, se quisermos o resultado da integral com uma determinada precisão, começamos calculando a integral com 2 pontos, vamos aumentando o número de pontos e comparando os resultados obtidos. Quando 2 resultados consecutivos tiverem o mesmo número de casas decimais iguais teremos o resultado com a precisão desejada. O próximo exemplo ilustra este fato.

Exemplo 11.17 - *Calcular*

$$\int_0^{1.2} e^x \; \cos \; x \; dx \; ,$$

com 3 casas decimais corretas, usando fórmula de quadratura de Gauss.

Solução: Como o intervalo de integração é finito, mas não cincide com o intervalo $[-1,1]$, devemos fazer uma mudança de variável. Assim,

$$\left| \begin{array}{ccc} x & t & 1 \\ 0 & -1 & 1 \\ 1.2 & 1 & 1 \end{array} \right| \; = \; -2x \; + \; 1.2t \; + \; 1.2 \; = \; 0 \; .$$

Logo:

$$x \; = \; 0.6 \; (t+1) \quad \text{e} \quad dx \; = \; 0.6 \; dt \; .$$

Assim:

$$\int_0^{1.2} e^x \; \cos \; x \; dx \; = \; 0.6 \; \int_{-1}^{1} e^{0.6 \; (t+1)} \; \cos \; (0.6 \; (t+1)) \; dt \; = \; I \; ,$$

e portanto estamos nas condições da fórmula de Gauss-Legendre com $f(t) = e^{0.6 \; (t+1)} \; \cos \; (0.6 \; (t+1))$. Fixemos $\mathbf{n} = 1$. Pela Tabela 1, com $\mathbf{N} = 2$, temos que:

$$t_0 \; = \; -0.5774 \; , \qquad A_0 \; = \; A_1 \; = \; 1.0 \; ,$$

$$t_1 \; = \; 0.5774 \; .$$

Agora:

$$f(t_0) \; = \; f(-0.5774) \; = \; e^{0.2536} \; \cos(0.2536) \; = \; 1.2474 \; ,$$

$$f(t_1) \; = \; f(0.5774) \; = \; e^{0.9464} \; \cos(0.9464) \; = \; 1.5062 \; .$$

Portanto:

$$I \; = \; 0.6 \; \times (1.2474 + 1.5052) \; = \; 1.6522 \; .$$

Tomemos agora 3 pontos. Assim, pela Tabela 1, com $\mathbf{N} = 3$, temos que:

$$t_0 \; = \; -0.7746 \; , \quad A_0 \; = \; A_2 \; = \; 0.5556 \; ,$$

$$t_1 \; = \; 0 \; , \qquad A_1 \; = \; 0.8889 \; ,$$

$$t_2 \; = \; 0.7746 \; .$$

Agora:

$$\begin{aligned} f(t_0) &= f(-0.7746) = e^{0.1352} \cos(0.1352) = 1.1343 \ , \\ f(t_1) &= f(0) = e^{0.6} \cos(0.6) = 1.5038 \ , \\ f(t_2) &= f(0.7746) = e^{1.0648} \cos(1.0648) = 1.4058 \ . \end{aligned}$$

Portanto:

$$I = 0.6 \ (\times(1.1343 + 1.4058) \times 0.5556 + 1.5038 \times 0.8889)) = 1.6488 \ .$$

Comparando o resultado obtido com $\mathbf{n} = 1$ e $\mathbf{n} = 2$, vemos que só temos uma casa decimal correta. Tomemos então $\mathbf{n} = 3$. Assim, pela Tabela 1, com $\mathbf{N} = 4$, temos que:

$$\begin{aligned} t_0 &= -0.8611 \ , \quad A_0 = A_3 = 0.3479 \ , \\ t_1 &= -0.3400 \ , \quad A_1 = A_2 = 0.6521 \ , \\ t_2 &= 0.3400 \ , \\ t_3 &= 0.8611 \ . \end{aligned}$$

Agora:

$$\begin{aligned} f(t_0) &= f(-0.8611) = e^{0.08334} \cos(0.08334) = 1.0831 \ , \\ f(t_1) &= f(-0.3400) = e^{0.396} \cos(0.396) = 1.3709 \ , \\ f(t_2) &= f(0.3400) = e^{0.8040} \cos(0.8040) = 1.5503 \ , \\ f(t_3) &= f(0.8611) = e^{1.1167} \cos(1.1167) = 1.3400 \ . \end{aligned}$$

Portanto:

$$\begin{aligned} I &= 0.6 \ \times (\ (1.0831 + 1.3400) \times 0.3479 + (1.3709 + 1.5503) \times 0.6521) \\ &= 1.6487 \ . \end{aligned}$$

Comparando o resultado obtido com $\mathbf{n} = 2$ e $\mathbf{n} = 3$, vemos que temos três casas decimais iguais. Assim, temos obtido o resultado da integral com três casas decimais corretas. Observe que com este exemplo fica claro a superioridade das fórmulas de quadratura de Gauss, pois obtivemos a precisão desejada usando **3** pontos. Nos exemplos 11.2, 11.3 e 11.4 resolvemos esta integral, e para obter 3 casas decimais corretas usando as fórmulas de Simpson foram necessários **7** pontos e para a regra do trapézio seriam necessários **49** pontos, (ver exemplo 11.6).

Exercícios:

11.15 - *Calcular*

$$a) \quad \int_{-1}^{1} \left(z^3 + z^2 + z + 1\right) \ dz \ ,$$

$$b) \quad \int_{-2}^{0} \left(x^2 - 1\right) \ dx \ ,$$

por quadratura de Gauss e diretamente e comparar os resultados.

11.16 - *Usando quadratura de Gauss, calcular:*

$$\int_{-1}^{1} \left(1 - x^2\right)^{-1/2} \; x^2 \; dx \; .$$

11.17 - *Calcular exatamente, a menos de erros de arredondamento,*

$$\int_{-1}^{1} \left(\frac{1}{2+2x} \; + \; \frac{1}{2-2x} \right)^{1/2} \; dx \; .$$

11.18 - *Usando quadratura de Gauss, calcular:*

$$\int_{1}^{2} \frac{dx}{2(x-1)\sqrt{-x^2+3x-2}} \; ,$$

com 2 casas decimais corretas.

11.19 - *Calcular:*

$$\int_{0}^{1} \left(\frac{1}{4x} \; + \; \frac{1}{4-4x} \right)^{1/2} \; dx \; ,$$

com 2 casas decimais corretas, usando quadratura de Gauss.

11.20 - *Calcular:*

$$\int_{-\pi}^{\pi} \cos^4 \; \theta \; d\theta \; ,$$

usando a fórmula de Gauss-Tchebyshev.

11.21 - *Usando quadratura de Gauss, calcular:*

$$\int_{0}^{\infty} \left(\frac{x^3+4x+2}{e^{2x}} \right) \; e^x \; dx \; ,$$

exatamente, a menos de erros de arredondamento.

11.22 - *Calcular:*

$$\Gamma(\alpha) \; = \; \int_{0}^{\infty} \; e^{-x} \; x^{\alpha-1} \; dx, \text{ para } \alpha = 5 \; ,$$

usando quadratura de Gauss.

Observação: $\Gamma(\alpha)$ é conhecida como função gama e vale, para α inteiro, $\Gamma(\alpha+1) \; = \; \alpha!$

11.23 - *Calcular:*

$$\int_{1}^{\infty} \; \frac{e^{-t}}{t} \; dt \; ,$$

usando quadratura de Gauss sobre 4 pontos.

11.24 - *Calcular:*

$$\int_{-\infty}^{\infty} \; e^{-x^2} \; sen \; x \; dx \; ,$$

usando quadratura de Gauss sobre 3 pontos.

11.6 Exercícios Complementares

11.25 - *Obtenha a fórmula de integração de Newton-Cotes do tipo fechado, para integrar $f(x)$ com $n = 4$, ou seja sobre 5 pontos. Usando a fórmula obtida, aproxime:*

$$\int_2^3 x\ e^{\frac{x}{2}}\ dx,$$

sabendo que:

x	2	2.25	2.5	2.75	3.0
$e^{\frac{x}{2}}$	2.77	3.08	3.49	3.96	4.48

11.26 - *Calcule as integrais a seguir pela regra do trapézio e pelas regras $\frac{1}{3}$ e $\frac{3}{8}$ de Simpson usando 6 divisões do intervalo de integração. Compare os resultados.*

$$I) \int_1^{2.5} x\ lnx\ dx, \quad II) \int_{-1.5}^{0} x\ e^x\ dx\ .$$

11.27 - *Nas integrais do exercício anterior com quantas divisões do intervalo, podemos esperar obter erros menores que 10^{-5}?*

11.28 - *Considere os seguintes dados experimentais:*

x	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
$y(x)$	1.00	1.82	2.08	3.18	3.52	4.7	5.12	6.38	6.98	8.22	9.00

encontre a área sob a curva $y = y(x)$ usando: a regra $\frac{3}{8}$ de Simpson de 1.0 a 2.2 e a regra $\frac{1}{3}$ de Simpson de 2.2 a 3.0.

11.29 - *Considere a integral:*

$$I\ =\ \int_0^1 \frac{sen\ t}{t}\ dt\ =\ 0.45970.$$

mostre que o resultado obtido pela regra $\frac{1}{3}$ de Simpson com $h = 0.5$ é surpreendente próximo do resultado exato.Explique porque isso acontece. (Sugestão: analise o erro).

11.30 - *Em contraste com o exercício anterior, considere:*

$$I\ =\ \int_{0.1}^1 \frac{dx}{x}\ =\ 2.30259.$$

Usando a regra de Simpson sobre N intervalos obteve-se a tabela:

N	h	I	*Erro*
2	0.45	3.3500	−1.0474
4	0.225	2.2079	−0.1053
8	0.1125	2.3206	−0.0180

Explique a diferença entre estes resultados e o do exercício anterior.

11.31 - *Considere a função $f(x)$ dada pela tabela:*

x	−2	−1	0	1	2
$f(x)$	−1	5	1	5	35

a) *Avalie:*

$$I)\ \int_{-2}^{2} f(x)\ dx,$$

usando a fórmula $\frac{1}{3}$ *de Simpson.*

b) *Se os valores tabelados são de um polinômio de grau 3 o que pode ser afirmado sobre o erro cometido na aproximação de I) pela fórmula* $\frac{1}{3}$ *de Simpson?*

11.32 - *Considere a integral:*

$$\int_{0}^{1} (2x^3 - 3x^2 + 1)\ dx\ .$$

Justifique o porquê da regra do trapézio com $h = 1$ *ser exata para calcular tal integral.*

Sugestão: *Verifique que para* $0 < t < 0.5$ *vale:*

$$f(0.5 + t) - P_1(0.5 + t) = -(f(0.5 - t) - P_1(0.5 - t))$$

onde $f(x) = 2x^3 - 3x^2 + 1$ *e* $P_1(x)$ *é a reta que interpola* $f(x)$ *em* $x_0 = 0$ *e* $x_1 = 1$.

11.33 - *Aproxime pela regra de Simpson o comprimento de arco da curva:*

$$y = 4x^2 - 3x$$

de $(0, 0)$ *a* $(1, 1)$.

Obs: *Lembre que o comprimento de arco de uma curva* $(a, f(a))$ *a* $(b, f(b))$ *é dada por:*

$$\int_{a}^{b} \sqrt{1 + (y'(x))^2}\ dx.$$

11.34 - *Uma maneira de se obter numericamente valores da função* $f(x) = \ln x$ *é calcular numericamente valores da integral*

$$\ln x = \int_{1}^{x} \frac{dt}{t}\ .$$

Calcular:

i) $\ln 17$,

ii) $\ln 35$,

com erro relativo inferior a 10^{-4}.

11.35 - *Escolha uma regra de quadratura sobre pontos igualmente espaçados de h e avalie:*

$$\int_{-1}^{0} x\ e^x\ dx\ ,$$

com duas casas decimais corretas.

11.36 - *Considere a integral:*

$$I = \int_{0}^{0.8} (x^2 - \cos\ x)\ dx.$$

a) *Quantos intervalos seriam necessários para aproximar I usando a regra do trapézio, com erro inferior a* 10^{-2}.

b) *Calcule I com o h obtido no item* **a)**.

11.37 - *Suponha que se conhece o valor de uma função* $f(x)$ *através da seguinte tabela:*

x	1.0	1.2	1.4	1.7	2.0	2.3	2.65	3.0
$f(x)$	0.23	0.59	1.1	1.4	0.92	0.63	0.42	0.38

Como você procederia para calcular:

$$\int_{1.0}^{3.0} f(x)\ dx\ ,$$

com a maior precisão possível?

11.38 - *Considere a integral:*

$$I\ =\ \int_{\alpha}^{1} x^{-\frac{1}{2}}\ dx\ .$$

a) *Determinar o número de intervalos suficientes para garantir o cálculo de I, com 4 casas decimais corretas, usando a regra de Simpson, nos seguintes casos:*

i) $\alpha = 0.1$,

ii) $\alpha = 0.01$,

iii) $\alpha = 0$.

b) *Obter I com* $\alpha = 0$, *usando fórmula de quadratura de Gauss sobre 3 pontos.*

11.39 - *Uma maneira de avaliar integrais da forma:*

$$I\ =\ \int_{0}^{\infty} f(x)\ dx\ ,$$

é aproximar I, fazendo:

$$I\ \simeq\ I^*\ =\ \int_{0}^{k} f(x)\ dx\ ,$$

onde k é um inteiro escolhido de modo que, dado um valor $\delta > 0$, *vale* $|f(x)| < \delta$ *para todo* $x > k$.

Considere a integral:

$$I\ =\ \Gamma(m)\ =\ \int_{0}^{\infty} e^{-x}\ x^{m-1}\ dx\ .$$

a) *Verificar que, para* $\delta = 10^{-4}$ *e* $m = 3$, *para utilizarmos o procedimento acima devemos tomar* $k = 15$.

ii) *Em quantos subintervalos devemos dividir o intervalo* $[0, 15]$, *para obter* $I^* \simeq \Gamma(3)$ *com 4 casas decimais corretas, usando a regra do trapézio?*

iii) *Obter o valor exato de* $I = \Gamma(3)$, *usando fórmula de quadratura de Gauss adequada.*

11.40 - *Considere o problema: Calcular*

$$I = \int_{a}^{b}\int_{c}^{d} f(x,y)\ dy\ dx\ .$$

a) *Verifique que a aplicação da regra do trapézio primeiramente na direção Oy e depois na direção Ox, fornece:*

$$I \approx \frac{(b-a)}{2}\frac{(d-c)}{2}[f(a,c)+f(b,c)+f(a,d)+f(b,d)] \; .$$

b) *Verifique que discretizando [a,b] e [c,d] respectivamente pelos pontos:*

$$x_i = a + ih \; , \quad 0 \le i \le m \; , \quad h = \frac{b-a}{m} \; ;$$

$$y_j = a + jk \; , \quad 0 \le j \le n \; , \quad k = \frac{d-c}{n} \; ;$$

e então aplicando a regra do trapézio generalizada nas direções Oy e Ox, obtemos:

$$I \approx \frac{hk}{4}\sum_{i=0}^{m}\sum_{j=0}^{n} a_{ij} f(x_i, y_j) \; ,$$

onde:

$$a_{00} = a_{m0} = a_{0n} = a_{nn} = 1 \; ,$$
$$a_{i0} = a_{in} = 2 \; , \quad 1 \le i \le m-1 \; ;$$
$$a_{0j} = a_{mj} = 2 \; , \quad 1 \le j \le n-1 \; ;$$
$$a_{ij} = 4 \; , \quad 1 \le i \le m-1 \; , \quad 1 \le j \le n-1 \; .$$

c) *Usando a fórmula obtida em* **b)**, *com h = 0.5 e k = 0.25, avalie:*

$$\int_0^1 \int_0^{0.5} \sqrt{x^2+y^3} \; dy \; dx \; .$$

11.41 - *Considere a integral:*

$$I(a) \; = \; \int_0^a \; x^2 \; e^{-x} \; dx \; .$$

a) *Obtenha $I(1)$ com duas casas decimais corretas usando a regra de Simpson.*

b) *Calcule exatamente, a menos de erros de arredondamento $I(a), a \to \infty$, usando fórmula de quadratura de Gauss.*

11.42 - *Uma pessoa calculou:*

$$I \; = \; \int_{-1}^{1} \; (3x^3 + 2x^2 + 2) \; dx \; ,$$

usando a regra $\frac{1}{3}$ de Simpson sobre os pontos: -1, 0, 1 e outra pessoa calculou I usando a fórmula de Gauss-Legendre sobre os pontos: - 0.57735, 0.57735. Qual das duas obteve melhor resultado? (Suponha que em ambos os casos não houve erros de arredondamento).

11.43 - *Deseja-se calcular:*

$$\int_{-2}^{2} \frac{e^t}{\sqrt{2-t}\;\sqrt{2+t}} \; dt \; .$$

Se você só pode obter o valor do integrando em 2 pontos, quais deverão ser estes pontos, de modo que a resposta seja a mais exata possível?

11.44 - *Considere a tabela:*

x	0	0.5	1.0	1.5	2.0	2.5	3.0
$f(x)$	-2	-1.5	-0.5	1.5	4.5	9.0	17.0

Sabendo que a fórmula de quadratura:

$$\int_a^b f(x)\ dx = Af(w)\ , \quad a \leq w \leq b\ ,$$

é exata para polinômios de grau ≤ 1*, caclule* A *e* w *e use-os para aproximar:*

$$\int_0^3 f(x)\ dx.$$

11.45 - *Seja:*

$$\int_{-1}^2 f(x)\ dx \ = \ A_0\ f(x_0) + A_1\ f(x_1)\ ,$$

onde $x_0 = \dfrac{1+\sqrt{3}}{2}$ *e* $x_1 = \dfrac{1-\sqrt{3}}{2}$ *são os zeros do polinômio de grau 2 ortogonal em* $[-1, 2]$ *segundo a função peso* $\omega(x) = 1$.

a) *Calcule os coeficintes* A_0 *e* A_1 *tal que a fórmula seja exata para polinômios de grau* ≤ 3.

b) *Usando a fórmula de quadratura obtida em* **a)***, calcule:*

$$\int_{-2}^0 \frac{dx}{x+3} dx\ .$$

11.46 - *Determine uma fórmula de quadratura para aproximar:*

$$\int_0^1 x\ f(x) dx\ ,$$

que seja exata quando f(x) é um polinômio de grau ≤ 3*. Usando a fórmula obtida calcule:*

$$\int_0^1 (x^4 + x\ sen\ x)\ dx\ .$$

11.47 - *Considere a integral:*

$$I = \int_0^{1.6} x^{-x}\ dx\ .$$

Obtenha o valor aproximado de I, com 2 dígitos significativos corretos:

a) *usando fórmula de Simpson.*

b) *usando fórmula de quadratura de Gauss.*

Lembre-se que: $\lim_{x\to 0} x^x = 1$.

11.7 Problemas Aplicados e Projetos

11.1 - *Um corpo negro (radiador perfeito) emite energia em uma taxa proporcional à quarta potência de sua temperatura absoluta, de acordo com a equação de Stefan-Boltzmann,*

$$E = 36.9 \quad 10^{-12} \quad T^{-4} \ ,$$

onde E *= potência de emissão,* W/cm^2 *e* T *= temperatura,* K^o*.*

O que se deseja é determinar uma fração dessa energia contida no espectro visível, que é tomado aqui como sendo 4.10^{-5} *a* 7.10^{-5} *cm. Podemos obter a parte visível integrando a equação de Planck entre esses limites:*

$$E_{visível} = \int_{4.10^{-5}}^{7.10^{-5}} \frac{2.39 \quad 10^{-11}}{x^5(e^{1.432/Tx} - 1)} dx \ ,$$

onde x *= comprimento de onda, cm;* E *e* T *como definido acima.*

A eficiência luminosa é definida como a relação da energia no espectro visível para a energia total. Se multiplicarmos por 100 para obter a eficiência percentual e combinarmos as constantes, o problema torna-se o de calcular:

$$EFF = \left(64.77 \quad \int_{4.10^{-5}}^{7.10^{-5}} \frac{dx}{x^5(e^{1.432/Tx} - 1)} \right) / \ T^4 \ .$$

Obter a eficiência luminosa, com erro relativo $< 10^{-5}$*, nas seguintes condições:*

i) T_i *= 2000 ºK*
T_f *= 3000 ºK*
com incremento da temperatura igual a 250.

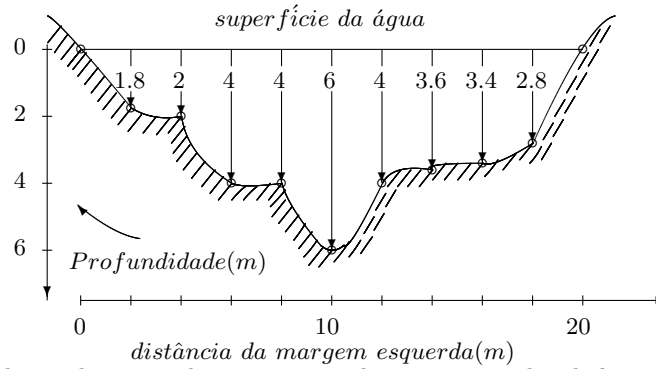
ii) T_i *= 2000 ºK*
T_f *= 3000 ºK*
com incremento da temperatura igual a 200.

onde T_i *e* T_f *são as temperaturas iniciais e finais, respectivamente.*

11.2 - *De um velocímetro de um automóvel foram obtidos as seguintes leituras de velocidade instantânea:*

$t(min.)$	0	5	10	15	20	25	30	35	40
$v(km/h)$	23	25	30	35	40	45	47	52	60

Calcule a distância em quilômetros, percorrida pelo automóvel usando a regra de Simpson.

11.3 - *A determinação da área da seção reta de rios e lagos é importante em projetos de prevenção de enchentes (para o cálculo de vazão da água) e nos projetos de reservatórios (para o cálculo do volume total de água). A menos que dispositivos tipo sonar sejam usados na obtenção do perfil do fundo de rios/lagos, o engenheiro civil deve trabalhar com valores da profundidade, obtidos em pontos discretos da superfície. Um exemplo típico de seção reta de um rio está mostrado na figura a seguir:*

Use uma fórmula de quadratura sobre pontos igualmente espaçados de h, para calcular a área da seção reta da figura dada acima.

11.4 - *A equação de Clapeyron encontrada no estudo das relações de propriedade termodinâmica pode ser expressa como:*

$$\frac{d \ln P}{dT} = \frac{\Delta H_r}{RT^2} \ , \tag{11.27}$$

onde P: pressão do vapor, T: temperatura absoluta, ΔH_r: entalpia da vaporização, R: constante de gás.

Esta temperatura, que é válida para um intervalo limitado de pressão e temperatura, pode ser usada para determinar a pressão de vapor em qualquer temperatura, reecrevendo-se (11.27) e integrando a partir de alguma pressão e temperatura conhecidas P_0, T_0. Mostre que fazendo isso obtemos:

$$\ln \frac{P}{P_0} = \int_{T_0}^{T} \frac{\Delta H_r}{RT^2} dT \ . \tag{11.28}$$

A solução de (11.28) requer o cálculo da integral indicada. Entretanto em muitos casos ΔH_r, não pode ser dada por uma expressão analítica conveniente e a integral deve então ser calculada por um método numérico.

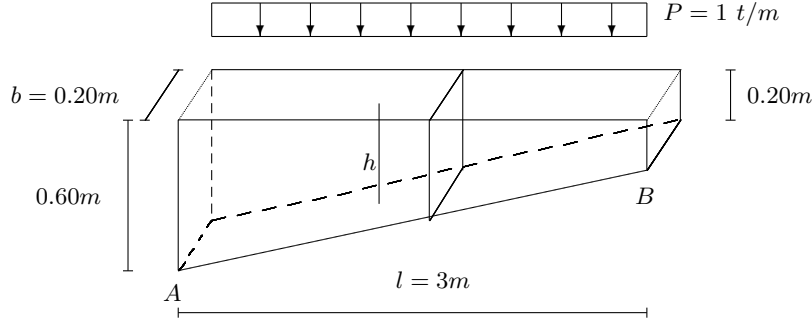
Considere uma substância para a qual os seguintes dados são conhecidos:

T	185	190	195	200	205	210
ΔH_r	81.307	80.472	79.568	78.714	77.859	77.002
	215	220	225	230	235	
	76.141	75.272	74.395	73.508	72.610	

R = 0.01614 K cal / Kg, P_0 = 0.028 atm em T_0 = 185 ^oK .

Determine a pressão do vapor a uma temperatura de 235 ^oK, usando 3, 5, 7, 9 e 11 pontos. Com 11 pontos é possível dizer quantas casas decimais estão corretas? Se a resposta for afirmativa diga qual é a precisão obtida.

Observe que nesse problema você não deve entrar com o valor de ϵ, mas sim comparar os resultados obtidos.

11.5 - *Seja a viga em balanço e o carregamento dado na figura a seguir:*

onde $E = 200t/cm^2$ *(módulo de elasticidade do concreto).*

O deslocamento vertical no ponto B pode ser obtido através da expressão:

$$\delta_{VB} = \int_A^B \frac{M_0 M_1}{EJ} dx \ ,$$

com

$$M_0 = \frac{1}{2}(\ell - x)^2, \quad M_1 = \ell - x, \quad J = \frac{bh^3}{12} \ ,$$

onde M_0 *é o momento da viga,* M_1 *é o momento da viga correspondente a uma carga unitária na direção e sentido do deslocamento e* J *é o momento de inércia de uma secção retangular de altura* h.

Determine o deslocamento vertical δ_{VB} *com erro relativo inferior a* 10^{-4}.

11.6 - *Na determinação da radiação luminosa emitida por um radiador perfeito é necessário calcular-se o valor da integral:*

$$Q = \int_{\lambda_1}^{\lambda_2} \frac{2\pi hc^2}{\lambda^5 (e^{\frac{hc}{k\lambda T}} - 1)} d\lambda,$$

onde:

Q = *radiação emitida por unidade de tempo por unidade de área entre os comprimentos de onda* λ_1 *e* λ_2*, em* $erg/cm^2.s$*,*

$\boldsymbol{\lambda_1}$ *e* $\boldsymbol{\lambda_2}$ = *limites inferior e superior, respectivamente, do comprimento de onda, em cm,*

h = *constante de Planck* = 6.6256×10^{-27} *erg.s,*

c = *velocidade da luz* = 2.99793×10^{10} *cm/s,*

k = *constante de Boltzmann* = 1.38054×10^{-16} *erg/k,*

T = *temperatura absoluta da superfície,* ^oK*,*

λ = *variável de integração = comprimento de onda, cm.*

Obter Q, com erro relativo $< 10^{-5}$*, nas seguintes condições:*

i) $\lambda_1 = 3.933666 \times 10^{-5}\ cm$,
$\lambda_2 = 5.895923 \times 10^{-5}\ cm$,
$T = 2000\ ^oK$.

ii) $\lambda_1 = 3.933666 \times 10^{-5}\ cm$,
$\lambda_2 = 5.895923 \times 10^{-5}\ cm$,
$T = 6000\ ^oK$.

11.7 - *A função de Debye é encontrada em termodinâmica estatística no cálculo do calor específico da água a volume constante de certas substâncias. A função é expressa por:*

$$D(x) = \frac{3}{x^3} \int_0^x \frac{y^3}{e^y - 1} dx \ .$$

Obter $D(x)$*, com erro relativo* $< 10^{-5}$*, nos seguintes casos:*

i) $x = 0.5$

ii) $x = 10$

iii) $x = 50$

11.8 - *Uma aproximação para a velocidade em função do tempo de um paraquedista em queda livre na atmosfera é dada pela equação:*

$$v(t) \ = \ \frac{gm}{c}\left(1 - e^{-\frac{c}{m}t}\right) \ ,$$

onde g *é a aceleração da gravidade* $(9.8m/s^2)$*,* m *é a massa do paraquedista* $(68kg)$*,* c *é o coeficiente de arrasto* $(12.5kg/s)$ *e* t *é o tempo (em s) a partir do ínicio da queda. Suponha que o paraquedista salte de uma altura de* $3000m$*. Sabendo que o espaço percorrido pelo paraquedista entre os instantes de tempo* a *e* b *é dado por:*

$$\Delta s \ = \ \int_a^b v(t)\ dt \ ,$$

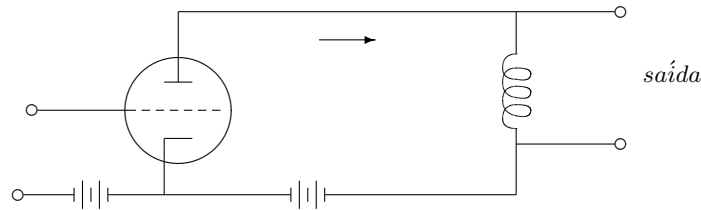
calcule a altura em que se encontra o paraquedista nos instante $t = 2s$ *e* $t = 10s$*. Em ambos os casos, utilize a regra* $\frac{1}{3}$ *de Simpson, com um número adequado de subintervalos para que o erro seja menor que* $1m$*.*

11.9 - *O etileno ocupa a quinta posição entre os produtos químicos mais fabricados nos Estados Unidos e o primeiro lugar entre os produtos químicos orgânicos, ao longo de um ano. Mais de 28 milhões de libras foram produzidas em 1985 e vendidas a U $ 22/libra. De todo etileno produzido,* 65% *é usado na fabricação de plásticos,* 20% *para óxido de etileno e etileno glicol,* 5% *para fibras e* 5% *para solventes.*

Deseja-se determinar o tamanho (volume) de um reator necessário para produzir 300 milhões de libras de etileno por ano do craqueamento de etano puro. A reação é irreversível e elementar. Além disso, deseja-se alcançar 80% *de conversão para o etano operando o reator isotermicamente a* $1100K$ *e à pressão de* $6atm$*. A equação para o reator é dada por:*

$$V \ = \ F_{A0} \int_0^x \frac{dx}{-\Gamma_A} \ ,$$

onde: V *é o volume do reator* (ft^3)*;* F_{A0} *é a taxa de alimentação do reagente (lb moles/s);* $-\gamma_A$ *é a taxa de reação* $(ft^3/lb\ mol)$*, e* x *é a conversão. A taxa de desaparecimento do etano* $(-\Gamma_A)$ *é dada por:* $-\Gamma_A = kC$ *, onde* k *é a constante de reação e* C*, a reconcentração do reagente (etano) é dada por:* $C = C_0(1-x)/(1+\epsilon)$*, com* C_0 *sendo a concentração inicial do reagente e* ϵ *o fator de mudança de volume. Usando uma regra de quadratura sobre pontos igualmente espaçados de* h*, determine o volume de um reator, dado que:* $F_{A0} = 0.425lb\ moles/s, k = 3.07s^{-1}, x = 0.8, C_0 = 0.00415lb\ moles/ft^3$ *e* $\epsilon = 1$*.*

11.10 - *A figura a seguir mostra um circuito típico contendo um amplificador.*

Muitos tipos de amplificadores são usados em instrumentos como transmissores de rádio e televisão, dispositivos de medidas, etc. Alguns tipos de amplificadores produzem correntes em pequeno pulso. Essa corrente é periódica no tempo, com T representando o período.

Para analisar o circuito, é usualmente necessário expressar a corrente em termos de uma função analítica. Usando a série de Fourrier truncada em m termos para I_p temos:

$$\begin{aligned} I_p(t) &= I_0 + I_1 \cos\left(\frac{2\pi t}{T}\right) + I_2 \cos\left(\frac{4\pi t}{T}\right) \\ &+ \ldots + I_m \cos\left(\frac{2m\pi t}{T}\right) = \sum_{k=0}^{m} I_k \cos\left(\frac{2k\pi t}{T}\right) , \end{aligned}$$

onde cada I_k é dado por:

$$I_k = \frac{2}{T} \int_0^T I_p(t) \cos\left(\frac{2k\pi t}{T}\right) dt , \quad k = 0, 1, \ldots, m.$$

Suponha que em certo experimento você mediu a corrente I_p em vários instantes de tempo e obteve a tabela a seguir:

$t(seg.)$	0	1	2	3	4	5	6	7	8	9	10
$I_p(t)$	100	94	80	60	31	0	−30	−58	−81	−95	−101
	11	12	13	14	15	16	17	18	19	20	
	−96	−82	−60	−30	0	32	59	80	95	99	

a) *Considerando $T = 20$, calcule: $I_0, I_1, \ldots, I_{20}$.*

b) *Desprezando os erros de arredondamento o que você pode concluir sobre a verdadeira expressão da função para $I_p(t)$?*

11.11 - *O serviço de proteção ao consumidor (SPC) tem recebido muitas reclamações quanto ao peso real do pacote de $5kg$ do açúcar vendido nos supermecados. Para verificar a validade das reclamações, o SPC contratou uma firma espacializada em estatística para fazer uma estimativa da quantidade de pacotes que realmente continham menos de $5kg$. Como é inviável a repezagem de todos os pacotes, a firma responsável pesou apenas uma amostra de 100 pacotes. A partir destes dados e utilizando métodos estatísticos eles puderam ter uma boa idéia do peso de todos os pacotes existentes no mercado.*

Chamando de x_i o peso do pacote i, tem-se que a média da amostra $\bar{x}$ é dada por:

$$\bar{x} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \;,$$

onde n é o número de pacotes da amostra. Serão omitidos os pesos, face ao elevado número de pacotes examinados.

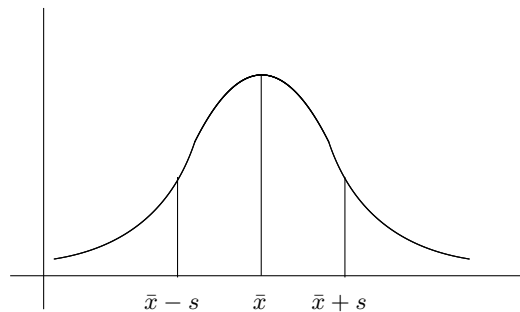
Calculando-se a média:

$$\bar{x} \;=\; \frac{1}{100} \;\times\; 499.1 \;=\; 4.991 kg \;.$$

O desvio padrão que é uma medida estatística que dá uma noção da dispersão dos pesos em relação à média é dado por:

$$S \;=\; +\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{x}-x_i)^2} \;.$$

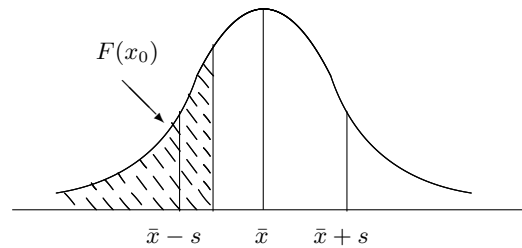
Para os dados deste problema tem-se que $S = 0.005 kg$.

Supondo-se verdadeira a hipótese de que a variação do peso dos pacotes não é tendenciosa, isto é, que o peso de um pacote é função de uma composição de efeitos de outras variáveis independentes, entre a quais podemos citar: regulagem da máquina de ensacar, variação da densidade do açúcar, leitura do peso, etc.; pode-se afirmar que a variável do peso tem uma distribuição normal. O gráfico da distribuição normal é apresentado a seguir:

A forma analítica desta função é:

$$f(x) \;=\; \frac{1}{S\sqrt{2\pi}} \; e^{\frac{1}{2}\left(\frac{x-\bar{x}}{S}\right)^2} \;.$$

O valor de $f(x)$ é a frequência de ocorrência do valor x. A integral de $f(x)$ fornece a frequência acumulada, isto é:

$$F(x_0) \;= \int_{-\infty}^{x_0} f(x)\; dx \;,$$

é a probabilidade de que x assuma um valor menor ou igual a x_0. Graficamente, $F(x_0)$ é a área hachurada na figura a seguir:

No problema em questão, o que se deseja é determinar :

$$F(5) \;=\; \int_{-\infty}^{5} \frac{1}{0.005\,\sqrt{2\pi}}\; e^{\frac{1}{2}\left(\frac{x-4.991}{0.005}\right)^2} \; dx \; .$$

Observações:

1) $\int_{-\infty}^{\infty} f(x)\; dx = 1.$

2) *A curva é simétrica em relação a média* ($\bar{x}$), *logo:*

$$\int_{-\infty}^{\bar{x}} f(x)\; dx \;=\; \int_{\bar{x}}^{\infty} f(x)\; dx \;=\; 0.5.$$

11.12 *-A definição da integral imprópria é:*

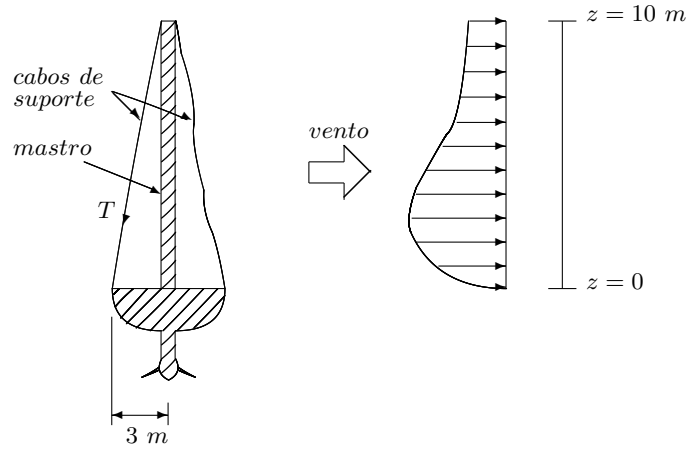
$$I \;=\; \int_{a}^{\infty} f(x)\; dx = \lim_{b \to \infty} \int_{a}^{b} f(x)\; dx \; .$$

Se essa integral for convergente podemos avaliá-la aproximadamente por método numérico. Por exemplo, a integral exponencial $Ei(x)$ *pode ser avaliada tomando o limite superior* U *suficientemente grande em:*

$$Ei(x) \;=\; \int_{x}^{\infty} \frac{e^{-v}}{v}\; dv \;\simeq\; \int_{x}^{U} \frac{e^{-v}}{v}\; dv \; .$$

Sabemos que U *é "suficientemente grande" quando as contribuições adicionais ao fazer* U *maior são desprezíveis. Estimar* $EI(0.5)$ *.*

Note que pode-se usar subintervalos maiores a medida que v *cresce. Compare o valor obtido com o valor tabular: 0.5598.*

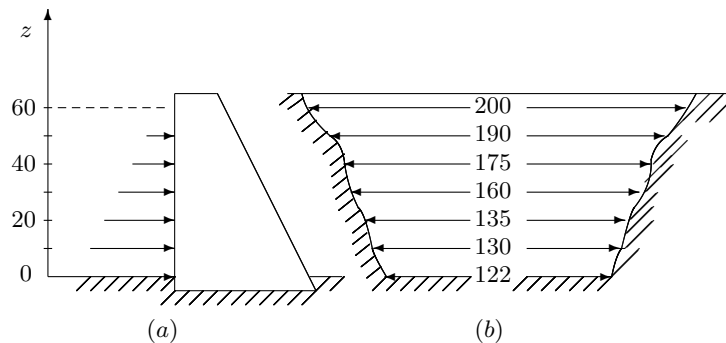
11.13 *- A seção reta de um veleiro está mostrada na figura a seguir:*

A força que o vento exerce sobre o mastro (devido às velas), varia conforme a altura z (em metros) a partir do convés. Medidas experimentais constataram que a força resultante exercida sobre o mastro (em N) é dada pela equação:

$$F \;=\; \int_0^{10} \; f(z)\; dz \;, \qquad f(z) \;=\; \frac{z}{4+z}\; e^{\frac{-2z}{10}} \;.$$

Deseja-se saber a linha de ação de F, isto é, o ponto onde pode-se aplicar uma força de mesmo módulo, direção e sentido de F, tal que o efeito sobre o mastro seja o mesmo de F. Esse ponto, localizado a uma altura d do convés do barco, pode ser determinado a partir da seguinte equação:

$$d \;=\; \frac{\int_0^{10} \; z\; f(z)\; dz}{\int_0^{10} \; f(z)\; dz} \;.$$

Pede-se então calcular o valor de d, usando fórmula de quadratura sobre pontos igualmente espaçados de h.

11.14 - *Suponha que a água em uma represa exerce uma pressão sobre a face esquerda da mesma, como mostrada na figura:*

Essa pressão pode ser caracterizada pela expressão:

$$p(z) \;=\; \rho g(D-z) \;,$$

onde $p(z)$ é a pressão (em N/m^2) na altura z (em m) a partir do fundo do represa. A densidade da água ρ é suposta constante e vale 10^3 kg/m^3, a aceleração da gravidade vale $9.8m/s^2$, e D é a altura (em m) da superfície da água a partir do fundo do represa. Sabe-se que a pressão aumenta linearmente com a profundidade, como mostrado em (a). A força total f_t sobre a face esquerda da represa pode ser calculada multiplicando-se a pressão pela área da face da represa. A largura da represa para diferentes profundidades, está mostrada em (b). Assuma que a largura da represa varia linearmente desde $200m$ *(na superfície) até* $122m$ *(a 60 m de profundidade).*

Assim a força resultante sobre a face da represa pode ser obtida através de:

$$f_t \; = \; \int_0^D \; \rho \; g \; \omega(z) \; (D - z) \; dz \; ,$$

onde $\omega(z)$ é a largura da represa na altura z a partir do fundo. Determine a altura d da linha de ação da força resultante, que pode ser obtida através do cálculo de:

$$d \; = \; \frac{\int_0^D \; z \; \rho \; g \; \omega(z) \; (D - z)}{f_t} \; dz \; .$$

TABELA 1 $\int_{-1}^{1} f(x)dx$

x_i		A_i
	N = 2	
0.5773502691		(1)0.1000000000
	N = 3	
0.7745966692		0.5555555555
0.0000000000		0.8888888888
	N = 4	
0.8611363115		0.3478548451
0.3399810435		0.6521451548
	N = 5	
0.9061798459		0.2369268850
0.5384693101		0.4786286704
0.0000000000		0.5688888888
	N = 6	
0.9324695142		0.1713244923
0.6612093864		0.3607615730
0.2386191860		0.4679139345
	N = 7	
0.9491079123		0.1294849661
0.7415311855		0.2797053914
0.4058451513		0.3818300505
0.0000000000		0.4179591836
	N = 8	
0.9602898564		0.1012285362
0.7966664774		0.2223810344
0.5255324099		0.3137066458
0.1834346424		0.3626837833

TABELA 2 $\int_{-1}^{1} \left(1-x^2\right)^a f(x)dx$

x_i		A_i
	$a=-1/2$	
	N = 2	
0.7071067811		(1)0.1570796326
	N = 3	
0.8660254037		(1)0.1047197551
0.0000000000		(1)0.1047197551
	N = 4	
0.9238795325		0.7853981633
0.3826834323		0.7853981633
	N = 5	
0.9510565162		0.6283185307
0.5877852522		0.6283185307
0.0000000000		0.6283185307
	N = 6	
0.9659258262		0.5235987755
0.7071067811		0.5235987755
0.2588190451		0.5235987755
	N = 7	
0.9749279121		0.4487989505
0.7818314824		0.4487989505
0.4338837391		0.4487989505
0.0000000000		0.4487989505
	N = 8	
0.9807852804		0.3926990816
0.8314696123		0.3926990816
0.5555702330		0.3926990816
0.1950903220		0.3926990812

TABELA 3 $\int_0^\infty e^{-x} f(x)dx$

x_i		A_i
	N = 2	
0.5857864376		0.8535533905
(1)0.3414213562		0.1464466094
	N = 3	
0.4157745567		0.7110930099
(1)0.2294280360		0.2785177335
(1)0.6289945082		(-1)0.1038925650
	N = 4	
0.3225476896		0.6031541043
(1)0.1745761101		0.3574186924
(1)0.4536620296		(-1)0.3888790851
(1)0.9395070912		(-3)0.5392947055
	N = 5	
0.2635603197		0.5217556105
(1)0.1413403059		0.3986668110
(1)0.3596425771		(-1)0.7594244968
(1)0.7085810005		(-2)0.3611758679
(2)0.1264080084		(-4)0.2336997238
	N = 6	
0.2228466041		0.4589646739
(1)0.1188932101		0.4170008307
(1)0.2992736326		0.1133733820
(1)0.5775143569		(-1)0.1039919745
(1)0.9837467418		(-3)0.2610172028
(2)0.1598297398		(-6)0.8985479064

TABELA 4 $\int_{-\infty}^{\infty} e^{-x^2} f(x)dx$

x_i		A_i
	N = 2	
0.7071067811		0.8862269254
	N = 3	
(1)0.1224744871		0.2954089751
0.0000000000		(1)0.1181635900
	N = 4	
(1)0.1650680123		(-1)0.8131283544
0.5246476323		0.8049140900
	N = 5	
(1)0.2020182870		(-1)0.1995324205
0.9585724646		0.3936193231
0.0000000000		0.6453087204
	N = 6	
(1)0.2350604973		(-2)0.4530009905
(1)0.1335849074		0.1570673203
0.4360774119		0.7246295952
	N = 7	
(1)0.2651961356		(-3)0.9717812450
(1)0.1673551628		(-1)0.5451558281
0.8162878828		0.4256072526
0.0000000000		0.8102646175
	N = 8	
(1)0.2930637420		(-3)0.1996040722
(1)0.1981656756		(-1)0.1707798300
(1)0.1157193712		0.2078023258
0.3811869902		0.6611470125

Capítulo 12

Solução Numérica de Equações Diferenciais Ordinárias

12.1 Introdução

Muitos problemas encontrados em engenharia e outras ciências podem ser formulados em termos de equações diferenciais. Por exemplo, trajetórias balísticas, teoria dos satélites artificiais, estudo de redes elétricas, curvaturas de vigas, estabilidade de aviões, teoria das vibrações, reações químicas e outras aplicações estão relacionadas com equações diferenciais. O objetivo deste capítulo é apresentar uma introdução à resolução de equações diferenciais ordinárias através de métodos numéricos.

A equação:

$$y' = f(x,y) \, , \tag{12.1}$$

é chamada **Equação Diferencial de primeira ordem**. Nesta equação f é uma função real dada, de duas variáveis reais x e y, e y é uma função incógnita da variável independente x. Além disso, y e f podem ser vetores, caso em que teremos um sistema de equações diferenciais de primeira ordem. Trataremos inicialmente apenas o caso escalar.

Resolver (12.1) corresponde a se determinar uma função $y = y(x)$, diferenciável, com $x \in [a, b]$ tal que $y'(x) = f(x, y(x))$. Qualquer função que satisfaça essa propriedade é uma solução da equação diferencial (12.1). Por exemplo, a função $y(x) = Ce^x$ é, para qualquer valor da constante C, uma solução da equação diferencial $y' = y$. Assim, cada equação diferencial de primeira ordem possui um número infinito de soluções. Contudo, podemos selecionar uma solução particular, se junto com a equação diferencial for dado o valor da solução $y(x)$ em um ponto, por exemplo, $y(x_0) = y_0$ (chamada condição inicial). Se para a equação diferencial: $y' = y$ é dado que $y(0) = 1$ então obtemos $C = 1$ e assim a solução do **(p.v.i.)** é $y(x) = e^x$.

A equação diferencial juntamente com a condição inicial constituem um **problema de valor inicial**, **(p.v.i)**, isto é:

$$\begin{cases} y' = f(x,y) \\ y(x_0) = y_0 \end{cases} \tag{12.2}$$

O seguinte teorema, estabelece condições sobre a $f(x, y)$ as quais garantem a existência de uma única solução do **(p.v.i)** (12.2).

Teorema 12.1 - Existência e Unicidade - *Seja $f(x,y)$ definida e contínua em* $D = \{(x,y)\ /\ a \leq x \leq b\ ;\ -\infty < y < \infty\ ;\ a$ e b finitos$\}$ *. Suponhamos que existe constante $L > 0$ tal que*

$$|f(x,y) - f(x,y^*)| \leq L|y - y^*|\ ,\ \forall (x,y)\ ,\ (x,y^*) \in D\ .$$

Então, se y_0 é um número dado, existe uma única solução $y(x)$ do **(p.v.i.)** *(12.2), onde $y(x)$ é contínua e diferenciável para todo $(x,y) \in D$.*

Prova: A prova deste teorema pode ser encontrada em [Henrice, 1962].

A grande maioria das equações encontradas na prática não podem ser solucionadas analiticamente; o recurso de que dispomos é o emprego de métodos numéricos. Para todos os métodos descritos neste capítulo consideraremos o **(p.v.i.)** (12.2) com a hipótese que f satisfaz as condições do Teorema (12.1), o qual garante que o problema tem uma única solução continuamente diferenciável, a qual indicaremos por $y(x)$. Também consideraremos a sequência de pontos $\{x_n\}$ definida por:

$$x_n = x_0 + nh\ ;\ n = 0, 1, \ldots, N\ ,$$

onde $x_0 = a,\ x_N = b$, e $N = \dfrac{b-a}{h}$.

Dizemos que o comprimento do intervalo, h, é o **tamanho do passo**, os pontos x_n são os **pontos da malha** e N é o **número de passos**.

Uma propriedade importante dos métodos computacionais para a solução de (12.2) é a **discretização**, isto é, desejamos obter a solução aproximada do **(p.v.i.)** não num intervalo contínuo $a \leq x \leq b$, mas sim num conjunto discreto de pontos $\{x_n\ /\ n = 0, 1, \ldots, N\}$.

Denotaremos por: y_n uma aproximação para a solução teórica em x_n, isto é: $y_n \simeq y(x_n)$ e por $f_n = f(x_n, y_n)$.

Nosso objetivo é então determinarmos aproximações y_n da solução verdadeira $y(x_n)$ nos pontos da malha. Portanto a solução numérica será uma tabela de valores dos pares $(x_n, y_n), y_n \simeq y(x_n)$.

Veremos a seguir alguns métodos para solução numérica de (12.2).

12.2 Método de Taylor de Ordem q

O primeiro método que discutiremos não é exatamente um método numérico, mas é algumas vezes usado em combinação com esquemas numéricos; é de aplicabilidade quase que geral, e servirá como introdução para outras técnicas que estudaremos.

Consideremos o **(p.v.i.)** (12.2). A função f pode ser linear ou não, mas vamos admitir que f seja contínua e suficientemente derivável em relação a x e y. Seja $y(x)$ é a solução exata de (12.2). A expansão em série de Taylor para $y(x_n + h)$ em torno do ponto x_n, é dada por:

$$\begin{aligned} y(x_n + h) &= y(x_n) + hy'(x_n) + \tfrac{h^2}{2!}y''(x_n) + \ldots \\ &+ \tfrac{h^q}{q!}y^{(q)}(x_n) + \tfrac{h^{q+1}}{(q+1)!}y^{(q+1)}(\xi_n)\ , \quad x_n < \xi_n < x_n + h\ , \end{aligned} \tag{12.3}$$

onde o último termo é o **erro de truncamento local**.

As derivadas na expansão (12.3) não são conhecidas explicitamente, uma vez que a solução exata não é conhecida. Contudo, se f é suficientemente derivável, elas podem ser obtidas considerando-se a derivada total de $y' = f(x,y)$ com respeito a x, tendo em mente que f é uma função implícita de y.

Assim sendo, obtemos para as primeira derivadas:

$$\begin{aligned}
y' &= f(x,y) \ , \\
y'' &= f' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\,\frac{dy}{dx} = f_x + f_y\, f \ , \\
y''' &= f'' = \frac{\partial f_x}{\partial x} + \frac{\partial f_x}{\partial y}\,\frac{dy}{dx} + \left[\frac{\partial f_y}{\partial x} + \frac{\partial f_y}{\partial y}\,\frac{dy}{dx}\right] f \\
&+ f_y \left[\frac{\partial f}{\partial x} + \frac{\partial f}{\partial y}\,\frac{dy}{dx}\right] \\
&= f_{xx} + f_{xy}\, f + f_{yx}\, f + f_{yy} f^2 + f_y f_x + f_y^2\, f \\
&= f_{xx} + 2 f_{xy}\, f + f_{yy} f^2 + f_x\, f_y + f_y^2\, f \ . \\
&\vdots
\end{aligned} \tag{12.4}$$

Continuando desta maneira, podemos expressar qualquer derivada de y em termos de $f(x,y)$ e de suas derivadas parciais. Contudo, é claro que, a menos que $f(x,y)$ seja uma função muito simples, as derivadas totais de ordem mais elevada tornam-se cada vez mais complexas. Por razões práticas deve-se, então, limitar o número de termos na expansão (12.3). Truncando então a expansão (12.3), após $(q+1)$ termos; obtemos:

$$y(x_n + h) = y(x_n) + h\, f(x_n, y(x_n)) + \ldots + \frac{h^q}{q!}\, f^{(q-1)}(x_n, y(x_n)) \ .$$

Esta equação pode ser interpretada como uma relação aproximada entre valores exatos da solução de (12.2). Uma relação exata entre valores aproximados da solução de (12.2) pode ser obtida substituindo-se $y(x_n)$ por y_n e $f^{(j)}(x_n, y(x_n))$ por $f_n^{(j)}; j = 0, 1, \ldots, q-1$. Fazendo isso, obtemos:

$$y_{n+1} = y_n + h\, f_n + \frac{h^2}{2!}\, f'_n + \ldots + \frac{h^q}{q!}\, f_n^{(q-1)} \ ; \tag{12.5}$$

que é chamado **Método de Taylor de ordem** q.

Exemplo 12.1 - *Resolver o* **(p.v.i)**:

$$\begin{cases} y' = -y + x + 2 \quad ; \\ y(0) = 2 \qquad\qquad 0 \le x \le 0.3 \quad ; \quad h = 0.1 \ , \end{cases}$$

usando o método de Taylor de ordem 3.

Solução: Temos que o método de Taylor de ordem 3 é dado por (12.5), com $q = 3$, isto é:

$$y_{n+1} = y_n + h\, f_n + \frac{h^2}{2}\, f'_n + \frac{h^3}{3!}\, f''_n \ . \tag{12.6}$$

Desde que $f = -y + x + 2$, obtemos:

$$\begin{aligned}
f' &= 1 + (-1)(-y + x + 2) = y - x - 1 \ , \\
f'' &= -1 + (1)(-y + x + 2) = -y + x + 1 \ .
\end{aligned}$$

Fazendo $n = 0$ em (12.6), obtemos:

$$y_1 \;=\; y_0 + h\; f_0 + \frac{h^2}{2}\; f_0' + \frac{h^3}{3!}\; f_0'' \; .$$

Observe que para calcular y_1 devemos avaliar f e suas derivadas em (x_0, y_0). Temos que o intervalo, onde desejamos obter a solução do **(p.v.i.)**, é $[0, 0.3]$ e $y_0 = 2$. Assim $(x_0, y_0) = (0, 2)$. Logo:

$$\begin{aligned}
f_0 &= f(x_0, y_0) = f(0, 2) = -2 + 0 + 2 = 0 \; , \\
f_0' &= f'(x_0, y_0) = f'(0, 2) = 2 - 0 - 1 = 1 \; , \\
f_0'' &= f''(x_0, y_0) = f''(0, 2) = -2 + 0 + 1 = -1 \; .
\end{aligned}$$

Portanto:

$$\begin{aligned}
y_1 &= 2 \;+\; 0.1(0) + \frac{(0.1)^2}{2}(1) + \frac{(0.1)^3}{3!}(-1) \\
&= 2.0048 \;\simeq y(x_1) \;=\; y(0.1) \; .
\end{aligned}$$

Fazendo agora $n = 1$ em (12.6), obtemos:

$$y_2 \;=\; y_1 + h\; f_1 + \frac{h^2}{2}\; f_1' + \frac{h^3}{3!}\; f_1'' \; .$$

Assim para calcular y_2 devemos avaliar a f e suas derivadas em (x_1, y_1). Temos que $h = 0.1 \;\rightarrow$ $(x_1, y_1) = (0.1, 2.0048)$. Portanto:

$$\begin{aligned}
f_1 &= f(x_1, y_1) = f(0.1, 2.0048) = -2.0048 + 0.1 + 2 = 0.0952 \; , \\
f_1' &= f'(x_1, y_1) = f'(0.1, 2.0048) = 2.0048 - (0.1) - 1 = 0.9048 \; , \\
f_1'' &= f''(x_1, y_1) = f''(0.1, 2.0048) = -2.0048 + 0.1 + 1 = -0.9048 \; .
\end{aligned}$$

Logo:

$$\begin{aligned}
y_2 &= 2.0048 + 0.1(0.0952) + \frac{(0.1)^2}{2}(0.9048) + \frac{(0.1)^3}{3!}(-0.9048) \\
&= 2.0186 \;\simeq y(x_2) \;=\; y(0.2) \; .
\end{aligned}$$

Finalmente, fazendo $n = 2$ em (12.6), obtemos:

$$y_3 \;=\; y_2 + h\; f_2 + \frac{h^2}{2}\; f_2' + \frac{h^3}{3!}\; f_2'' \; .$$

Temos que $(x_2, y_2) = (0.2, 2.0186)$. Assim:

$$\begin{aligned}
f_2 &= f(x_2, y_2) = f(0.2, 1.9603) = -2.0186 + 0.2 + 2 = 0.1814 \; , \\
f_2' &= f'(x_2, y_2) = f'(0.2, 1.9603) = 2.0186 - 0.2 - 1 = 0.8186 \; , \\
f_2'' &= f''(x_2, y_2) = f''(0.2, 1.9603) = -2.0186 + 0.2 + 1 = -0.8186 \; .
\end{aligned}$$

Portanto:

$$
\begin{aligned}
y_3 &= 2.0186 + 0.1(0.1814) + \frac{(0.1)^2}{2}(0.8186) + \frac{(0.1)^3}{3!}(-0.8186) \\
&= 2.0406 \simeq y(x_3) = y(0.3) .
\end{aligned}
$$

Assim a solução do **(p.v.i.)** dado é:

x_n	y_n	$y(x_n)$
0	2	2
0.1	2.0048	2.00484
0.2	2.0186	2.01873
0.3	2.0406	2.04082

onde a última coluna foi obtida através da solução exata do **(p.v.i.)** que é: $y(x) = e^{-x} + x + 1$.

Observe que nem sempre podemos aplicar o método de Taylor de ordem q, com q qualquer, como pode ser observado no próximo exemplo.

Exemplo 12.2 - *Resolver o* **(p.v.i)***:*

$$
\begin{cases}
y' = y^{\frac{1}{3}} \\
y(0) = 0 \; ; \quad x \in [0, 0.3] \; ; \quad h = 0.1.
\end{cases}
$$

usando o método de Taylor de ordem 3.

Solução: Temos que o método de Taylor de ordem 3 é dado por (12.6). Inicialmente calculamos as derivadas da f. Temos que:

$$
\begin{aligned}
y' &= f \\
y'' &= f' = \frac{1}{3} y^{-\frac{2}{3}} y^{\frac{1}{3}} = \frac{1}{3} y^{-\frac{1}{3}} ,
\end{aligned}
$$

Observe que a derivada segunda de y não existe, desde que $y_0 = 0$. Assim, só podemos utilizar o método de Taylor com $q = 1$.

Assim as desvantagens do método de Taylor de ordem q são: calcular as derivadas da f e a cada passo avaliá-las nos pontos (x_n, y_n), e o fato de nem sempre podermos utilizar o método com q qualquer.

12.3 Métodos Lineares de Passo Múltiplo

Inicialmente, para resolver o **(p.v.i.)** (12.2), descreveremos os **Métodos Lineares de Passo Múltiplo** ou **Métodos de k-passos**.

Definição 12.1 - *Um* **método linear de passo múltiplo** *é definido pela seguinte relação:*

$$\sum_{j=0}^{k} \alpha_j \; y_{n+j} \;=\; h \sum_{j=0}^{k} \beta_j \; f_{n+j} \;, \tag{12.7}$$

onde α_j *e* β_j *são constantes arbitrárias independentes de* n*, com* $\alpha_k \neq 0$ *e* α_0 *e* β_0 *não ambos nulos. Vamos supor* $\alpha_k = 1$.

Dizemos que o método (12.7) é **explícito** se $\beta_k = 0$ e **implícito** se $\beta_k \neq 0$.
Os métodos de passo múltiplo podem ser obtidos de várias maneiras. Veremos a seguir algumas técnicas de como obter tais métodos .

12.3.1 Obtidos do Desenvolvimento de Taylor

Descreveremos aqui como obter métodos lineares de passo múltiplo para resolver (12.2), baseados no desenvonvimento da solução exata do **(p.v.i.)** em série de Taylor. Novamente, a função f pode ser linear ou não, mas vamos admitir que f seja contínua e suficientemente derivável em relação a x e y.

I) O método mais simples de passo múltiplo é obtido fazendo $q = 1$, no método de Taylor de ordem q, fórmula (12.5). Obtemos então o **método explícito de 1-passo**:

$$y_{n+1} \;=\; y_n + h\; f_n \;, \tag{12.8}$$

chamado **Método de Euler**.

Exemplo 12.3 - *Usando o método de Euler, resolver o* **(p.v.i.)** *do exemplo 12.1.*

Solução: Fazendo $n = 0$, em (12.8), obtemos:

$$y_1 = y_0 + h\; f_0 = 2 \;+\; 0.1(0) = 2 \;\simeq\; y(x_1) = y(0.1) \;,$$

onde f_0 foi calculada no exemplo 12.1. Fazendo agora $n = 1$, em (12.8), segue que:

$$y_2 = y_1 + hf_1 = 2 + 0.1(0.1) \;=\; 2.01 \;\simeq\; y_(x_2) = y(0.2) \;,$$

desde que $f_1 = f(x_1, y_1) = f(0.1, 2) = -2 + 0.1 + 2 = 0.1$. Finalmente, fazendo $n = 2$ em (12.8), obtemos:

$$y_3 = y_2 + hf_2 = 2.01 + 0.1(0.19) \;=\; 2.019 \;\simeq y(x_3) \;=\; y(0.3) \;,$$

desde que $f_2 = f(x_2, y_2) = f(0.2, 2.01) = -2.01 + 0.2 + 2 = 0.19$.

Assim a solução do **(p.v.i.)** dado é:

x_n	y_n
0	2
0.1	2
0.2	2.01
0.3	2.019

Compare os resultados obtidos com os do exemplo 12.1. Como pode ser observado os resultados obtidos pelo método de Euler não são de boa qualidade. Em geral, o método de Euler tem, na verdade, mais importância teórica do que prática. É necessário então estabelecermos métodos mais precisos.

II) Considere agora o desenvolvimento de $y(x_n + h)$ e $y(x_n - h)$ em série de Taylor em torno do ponto x_n, isto é:

$$y(x_n + h) = y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(x_n) + \frac{h^3}{3!}y'''(x_n) + \ldots ,$$

$$y(x_n - h) = y(x_n) - hy'(x_n) + \frac{h^2}{2!}y''(x_n) - \frac{h^3}{3!}y'''(x_n) + \ldots .$$

Calculando $y(x_n + h) - y(x_n - h)$, obtemos:

$$y(x_n + h) - y(x_n - h) = 2hy'(x_n) + \frac{h^3}{3}y'''(x_n) + \ldots .$$

Considerando apenas o primeiro termo do lado direito da expansão acima, substituindo $y(x_n + h)$ por y_{n+1}, $y(x_n - h)$ por y_{n-1} e $y'(x_n)$ por f_n, obtemos:

$$y_{n+1} - y_{n-1} = 2hf_n .$$

Esta fórmula pode ser colocada na forma (12.7) trocando n por $n + 1$. Fazendo isso, obtemos que:

$$y_{n+2} = y_n + 2hf_{n+1} , \qquad (12.9)$$

que é um **método explícito de 2-passos** chamado **Regra do Ponto Médio**.

Observe que para resolver um **(p.v.i.)** usando um método explícito de 2-passos, como é o caso da regra do ponto médio, devemos ter disponíveis além do valor de y_0, o valor de y_1. Assim, o valor de y_1 deve ser obtido de alguma outra forma, por exemplo, usando método numérico de 1-passo.

Exemplo 12.4 - *Resolver o* **(p.v.i.)** *do exemplo 12.1, através da regra do ponto médio. Use o método de Taylor de ordem 2, para obter os valores iniciais necessários.*

Solução: Temos que: o valor de y_0 é dado no **(p.v.i.)** e para calcular y_1, usaremos o método de Taylor de ordem 2, o qual é dado por (12.5), com $q = 2$, isto é:

$$y_{n+1} = y_n + h\, f_n + \frac{h^2}{2}\, f'_n .$$

Os valores de f_0 e de f'_0 já foram calculados no exemplo 12.1. Assim:

$$y_1 = 2 + (0.1)(0) + \frac{(0.1)^2}{2}(1) = 2.0050 \simeq y(x_1) = y(0.1) .$$

Agora $f_1 = f(x_1, y_1) = f(0.1, 2.0050) = -2.0050 + 0.1 + 2 = 0.0950$.

Assim, fazendo $n = 0$, em (12.9), obtemos que:

$$y_2 = y_0 + 2hf_1 = 2 + 2(0.1)(0.0950) = 2.0190 \simeq y(x_2) = y(0.2) .$$

Finalmente, desde que $f_2 = f(x_2, y_2) = f(0.2, 2.0190) = -2.0190 + 0.2 + 2 = 0.1810$, obtemos, fazendo $n = 1$, em (12.9), que:

$$y_3 = y_1 + 2hf_2 = 2.0050 + 2(0.1)(0.1810) = 2.0412 \simeq y(x_3) = y(0.3) .$$

Assim a solução do **(p.v.i.)** dado é:

x_n	y_n
0	2
0.1	2.0050
0.2	2.0190
0.3	2.0412

Observe que o resultado aqui é mais preciso do que o obtido pelo método de Euler.

12.3.2 Obtidos de Integração Numérica

Descreveremos aqui como obter métodos lineares de passo múltiplo para resolver o **(p.v.i.)**, (12.2), obtidos a partir de fórmulas de integração numérica.

Integrando a equação diferencial de primeira ordem, do **(p.v.i.)** (12.2), de x_n até x_{n+k}, obtemos:

$$\int_{x_n}^{x_{n+k}} y'(x)\ dx \ = \ \int_{x_n}^{x_{n+k}} f(x, y(x))\ dx\ . \tag{12.10}$$

Agora, desde que o lado esquerdo de (12.10) pode ser integrado exatamente, obtemos que a solução exata de (12.2) satisfaz a identidade:

$$y(x_{n+k}) - y(x_n) \ = \ \int_{x_n}^{x_n+k} f(x, y(x))\ dx\ , \tag{12.11}$$

para quaisquer dois pontos x_n e x_{n+k} em [a,b]. Assim, para diferentes valores de k, após aproximar a integral do lado direito de (12.11), obtemos diferentes métodos lineares de passo múltiplo. De fato:

I) Fazendo $k = 1$ em (12.11), obtemos:

$$y(x_{n+1}) - y(x_n) \ = \ \int_{x_n}^{x_{n+1}} f(x, y(x))\ dx\ ,$$

e assim podemos aplicar a regra do trapézio, fórmula (11.6), para calcular a integral, na expressão acima desde que a mesma está sendo avaliada entre dois pontos consecutivos. Fazendo isso, segue que:

$$y(x_{n+1}) = y(x_n) + \frac{h}{2}\left[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))\right]\ .$$

Substituindo $y(x_n)$ e $y(x_{n+1})$ por y_n e y_{n+1}, respectivamente, obtemos:

$$y_{n+1} \ = \ y_n + \frac{h}{2}\left[f_n + f_{n+1}\right]\ , \tag{12.12}$$

que é um **método implícito de 1-passo** chamado **método do trapézio**.

Observe que, (12.12) é uma equação implícita para y_{n+1}, uma vez que y_{n+1} aparece como argumento no segundo membro. Se $f(x, y)$ for uma função não linear, não teremos, em geral, condições de resolver (12.12) em relação a y_{n+1} de uma forma exata. Assim métodos implícitos serão usados nos métodos do tipo Previsor-Corretor que serão descritos mais adiante.

II) Fazendo $k = 2$ em (12.11), obtemos:

$$y(x_{n+2}) - y(x_n) \; = \; \int_{x_n}^{x_{n+2}} \; f(x, y(x)) \; dx \; ,$$

e assim podemos aplicar a regra $\frac{1}{3}$ de Simpson, fórmula (11.9), para calcular a integral na expressão acima desde que a mesma está sendo avaliada entre três pontos consecutivos. Fazendo isso, segue que:

$$y(x_{n+2}) = y(x_n) + \frac{h}{3} \left[f(x_n, y(x_n)) + 4 f(x_{n+1}, y(x_{n+1})) + f(x_{n+2}, y(x_{n+2})) \right] \; ,$$

e como no caso anterior, obtemos:

$$y_{n+2} \; = \; y_n + \frac{h}{3} \left[f_n + 4 f_{n+1} + f_{n+2} \right] \; , \tag{12.13}$$

que é um **método implícito de 2-passos** chamado **método de Simpson**.

Observe que para poder aplicar o método (12.13), precisamos além de utilizar métodos do tipo Previsor-Corretor, também obter valores iniciais por métodos de 1-passo.

Além de aproximar a integral do lado direito de (12.11), usando as fórmulas de Newton-Cotes do tipo fechado, dadas no Capítulo 11, podemos também obter métodos de k-passos, baseados em fórmulas de integração numérica, usando as fórmulas de Newton-Cotes do tipo aberto, como veremos a seguir.
III) Seja $P(x)$ o único polinômio de grau 2 passando pelos pontos:

$$(x_n, f_n) \; , \; (x_{n+1}, f_{n+1}) \; , \; (x_{n+2}, f_{n+2}) \; .$$

Usando a forma de Newton-Gregory para o polinômio de interpolação, fórmula (10.32), obtemos:

$$P(x) \; = \; f_n \; + \; (x - x_n) \Delta f_n \; + \; (x - x_n)(x - x_{n+1}) \; \frac{\Delta^2 f_n}{2!} \; .$$

Agora, desde que os pontos $x_i, \; i = n, n+1, n+2$ são igualmente espaçados de h, podemos fazer a seguinte mudança de variável: $u = \dfrac{x - x_n}{h}$, e assim:

$$P(x) \; = \; P(x_n + uh) \; = \; f_n \; + \; u \; \Delta f_n \; + \; \frac{u(u-1)}{2} \; \Delta^2 f_n \; .$$

Integrando a equação diferencial de primeira ordem, do **(p.v.i.)** (12.2), de x_{n+1} até x_{n+2}, substituindo $y(x_{n+2})$ e $y(x_{n+1})$ por y_{n+2} e y_{n+1}, respectivamente, e usando o fato que:

$$\int_{x_{n+1}}^{x_{n+2}} \; f(x, y(x)) \; dx \cong \int_{x_{n+1}}^{x_{n+2}} \; P(x) \; dx \; = \; \int_1^2 \; P(x_n + uh) \; h \; du \; ,$$

obtemos:

$$\begin{aligned} y_{n+2} \; - \; y_{n+1} \; &= \; h \int_1^2 \left[f_n \; + \; u \Delta f_n \; + \; \frac{u(u-1)}{2} \; \Delta^2 f_n \right] du \\ &= \; h \left[f_n u \; + \; \frac{u^2}{2} \Delta f_n \; + \; \frac{1}{2} \; \left(\frac{u^3}{3} - \frac{u^2}{2} \right) \Delta^2 f_n \right]_1^2 \; . \end{aligned}$$

Agora, pela fórmula (10.31), temos que as diferenças ordinárias de ordens 1 e 2, são dadas, respectivamente, por:

$$\begin{aligned}\Delta f_n &= f_{n+1} - f_n \ , \\ \Delta^2 f_n &= f_{n+2} - 2f_{n+1} + f_n \ .\end{aligned}$$

Assim, substituindo as diferenças ordinárias na expressão acima e agrupando os termos semelhantes, segue que:

$$y_{n+2} = y_{n+1} + \frac{h}{12} \ [-f_n + 8 \ f_{n+1} + 5f_{n+2}] \ , \tag{12.14}$$

que é um **método implícito de 2-passos** chamado **Método de Adams-Moulton**. Vale aqui a mesma observação dada no método de Simpson.

IV) De maneira semelhante ao método anterior, se aproximarmos $f(x, y(x))$ por um polinômio de interpolação sobre os pontos (x_n, f_n) , (x_{n+1}, f_{n+1}), isto é, por um polinômio do primeiro grau, e integrarmos a equação diferencial de primeira ordem, do **(p.v.i.)** (12.2), de x_{n+1} até x_{n+2}, obtemos:

$$y_{n+2} = y_{n+1} + \frac{h}{2} \ [-f_n + 3 \ f_{n+1}] \ , \tag{12.15}$$

que é um **método explícito de 2-passos** chamado **Método de Adams-Bashforth**. Como na regra do ponto médio, para aplicar este método devemos obter, inicialmente, o valor de y_1 por método de 1-passo.

Exemplo 12.5 - *Resolver o* **(p.v.i.)** *do exemplo 12.1, usando o método de Adams-Bashforth. Use o método de Taylor de ordem 2, para obter os valores iniciais necessários.*

Solução: Temos que: $y_0 = 2$ (condição inicial), $f_0 = 0$ e $y_1 = 2.005$, $f_1 = 0.095$) (calculado no exemplo 12.4). Assim, fazendo $n = 0$ em (12.15), segue que:

$$\begin{aligned}y_2 &= y_1 + \frac{h}{2}[-f_0 + 3f_1] = 2.005 + \frac{1}{2}[-0 + 3(0.095)] \\ &= 2.0193 \simeq y(x_2) = y(0.2) \ .\end{aligned}$$

Agora $f(x_2, y_2) = f(0.2, 2.0193) = -2.0193 + 0.2 + 2 = 0.1807$. Assim, fazendo $n = 1$ em (12.15), obtemos:

$$\begin{aligned}y_3 &= y_2 + \frac{h}{2}[-f_1 + 3f_2] = 2.0193 + \frac{1}{2}[-0.095 + 3(0.1807)] \\ &= 2.0417 \simeq y(x_3) = y(0.3) \ .\end{aligned}$$

Assim a solução do **(p.v.i.)** dado é:

x_n	y_n
0	2
0.1	2.005
0.2	2.0193
0.3	2.0417

Observe que todos os métodos de passo múltiplo obtidos via integração numérica satisfazem:

$$\alpha_k = 1 \ , \quad \alpha_j = -1 \quad \text{e} \quad \alpha_i = 0 \ , \ i = 0, 1, \ldots, j-1, j+1, \ldots, k-1 \ .$$

Existem outras maneiras de se obter métodos lineares de passo múltiplo, entretanto julgamos que os métodos aqui apresentados dão uma boa idéia ao leitor do que sejam tais métodos e como podem ser aplicados.

Exercícios

12.1 - *Mostre que, fazendo $k = 3$ em (12.11), e usando a fórmula (11.12), obtem-se:*

$$y_{n+3} \ = \ y_n + \frac{3h}{8} \left[f_n + 3(f_{n+1} + f_{n+2}) + f_{n+3} \right] \ ,$$

que é um **método implícito de 3-passos** *chamado* **método $\frac{3}{8}$ de Simpson**.

12.2 - *Considere os seguintes problemas de valor inicial:*

a) $$\begin{cases} y' \ = \ y^2 + 1 \\ y(0) \ = \ 0 \qquad ; \quad 0 \leq x \leq 1 \quad ; \quad h \ = \ 0.2 \end{cases}$$

b) $$\begin{cases} y' \ = \ -2xy \quad ; \\ y(0) \ = \ 1 \qquad 0 \leq x \leq 0.6 \quad ; \quad h \ = \ 0.3 \end{cases}$$

c) $$\begin{cases} y' \ = \ -xy \quad ; \\ y(0) \ = \ 2 \qquad 0 \leq x \leq 0.3 \quad ; \quad h \ = \ 0.1 \end{cases}$$

Resolva-os pelo:

a) *método de Euler,*

b) *método de Taylor de ordem 2,*

c) *regra do ponto médio,*

d) *método de Adams-Bashforth,*

usando para os itens **c)** *e* **d)**, *o item* **b)** *para obter os valores iniciais necessários,*

12.3.3 Ordem e Constante do Erro

Analisaremos aqui a **Ordem** e a **Constante do Erro**, para os métodos lineares de passo múltiplo definidos por (12.7).

Definição 12.2 - *Definimos o* **operador diferença linear** $\mathcal{L}$*, associado ao método linear de passo múltiplo (12.7), por:*

$$\mathcal{L}[y(x);h] \;=\; \sum_{j=0}^{k} \; [\alpha_j y(x+jh) \;-\; h\beta_j y'(x+jh)] \;, \tag{12.16}$$

onde $y(x)$ é uma função arbitrária continuamente diferenciável em $[a,b]$.

Expandindo $y(x+jh)$ e $y'(x+jh)$ em série de Taylor em torno do ponto x, desenvolvendo o somatório e agrupando os termos semelhantes, obtemos:

$$\mathcal{L}[y(x);h] \;=\; C_0 y(x) + C_1 h y'(x) + \ldots + C_q h^q y^{(q)}(x) + \ldots \;, \tag{12.17}$$

onde

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 + \ldots + \alpha_k \;, \\ C_1 &= \alpha_1 + 2\alpha_2 + \ldots + k\alpha_k \;-\; (\beta_0 + \beta_1 + \ldots + \beta_k) \;, \\ &\vdots \\ C_q &= \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \ldots + k^q\alpha_k) \;-\; \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \ldots + k^{q-1}\beta_k) \;. \end{aligned} \tag{12.18}$$

Definição 12.3 - *O operador diferença (12.16) e o método linear de passo múltiplo associado (12.7), têm* **ordem** *q, se em (12.17), $C_0 = C_1 = \ldots = C_q = 0$ e $C_{q+1} \;\neq\; 0$. C_{q+1} é chamada de* **constante do erro**.

Exemplo 12.6 - *Obter a ordem e a constante do erro para:*

a) *o método de Euler,*

b) *a regra do trapézio.*

Solução: Temos que o método de Euler é dado por (12.8), de onde deduzimos que:

$$\alpha_0 \;=\; -1 \;, \qquad \beta_0 \;=\; 1 \;,$$

$$\alpha_1 \;=\; 1 \;, \qquad \beta_1 \;=\; 0 \;.$$

Assim:

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 \;\Rightarrow\; C_0 = -1 + 1 = 0 \;, \\ C_1 &= \alpha_1 - (\beta_0 + \beta_1) \;\Rightarrow\; C_1 = 1 - (1 + 0) = 0 \;, \\ C_2 &= \frac{1}{2!}(\alpha_1) - (\beta_1) \;\Rightarrow\; C_2 = \frac{1}{2}(1) - (0) = \frac{1}{2} \;. \end{aligned}$$

Logo, $C_0 = C_1 = 0$ e $C_2 \neq 0$. Portanto a ordem do método de Euler é $q = 1$ e a constante do erro é $C_2 = \frac{1}{2}$.

O método do trapézio é dado por (12.12). Assim:

$$\alpha_0 = -1 \,, \qquad \beta_0 = \frac{1}{2} \,,$$

$$\alpha_1 = 1 \,, \qquad \beta_1 = \frac{1}{2} \,.$$

Portanto:

$$\begin{aligned}
C_0 &= \alpha_0 + \alpha_1 \;\Rightarrow\; C_0 = -1 + 1 = 0 \,, \\
C_1 &= \alpha_1 - (\beta_0 + \beta_1) \;\Rightarrow\; C_1 = 1 - (\frac{1}{2} + \frac{1}{2}) = 0 \,, \\
C_2 &= \frac{\alpha_1}{2!} - \beta_1 \;\Rightarrow\; C_2 = \frac{1}{2} - \frac{1}{2} = 0 \,, \\
C_3 &= \frac{\alpha_1}{3!} - \frac{\beta_1}{(2)!} \;\Rightarrow\; C_3 = \frac{1}{6} - \frac{1}{4} = -\frac{1}{12} \,.
\end{aligned}$$

Logo, $C_0 = C_1 = C_2 = 0$ e $C_3 \neq 0$. Portanto a ordem do método do trapézio é $q = 2$ e a constante do erro é $C_3 = -\frac{1}{12}$.

Exercício

12.3 - *Determinar a ordem e a constante do erro para:*

a) *regra do ponto médio,*

b) *método de Simpson,*

c) *método de Adams-Moulton,*

d) *método de Adams-Bashforth,*

e) *método $\frac{3}{8}$ de Simpson.*

12.3.4 Erro de Truncamento Local

Agora, podemos definir formalmente o erro de truncamento local de um método linear de passo múltiplo.

Definição 12.4 - *Definimos* **Erro de Truncamento Local**, *em x_{n+k} do método linear de passo múltiplo, definido por (12.7), por:*

$$T_{n+k} \;=\; \mathcal{L}\,[y(x_n);h] \;=\; \sum_{j=0}^{k} \,[\alpha_j y(x_{n+j}) \;-\; h\beta_j y'(x_{n+j})] \;\;,$$

onde $y(x)$ é a solução exata do **(p.v.i)** *(12.2).*

Observe que o erro de truncamento é chamado *local*, pois supomos que nenhum erro foi cometido anteriormente, isto é, impomos:

$$y_{n+j} \ = \ y(x_{n+j}) \ , \quad j = 0, 1, \ldots, k-1 \ ,$$

e então só consideramos o erro em y_{n+k}.

Pode-se mostrar que:

$$T_{n+k} \ = \ \left[1 - \beta_k \frac{\partial f}{\partial y}(x_{n+k}, \xi_{n+k})\right] (y(x_{n+k}) \ - \ y_{n+k}) \ . \qquad (12.19)$$

onde $\xi_{n+k} \ \in \ (y_{n+k}, y(x_{n+k}))$.

Supondo que a solução teórica $y(x)$ tem derivadas contínuas de ordem suficientemente elevadas, então para ambos, métodos implícitos e explícitos, de (12.19) pode ser deduzido que

$$y(x_{n+k}) \ - \ y_{n+k} \ = \ C_{q+1} h^{q+1} y^{(q+1)}(x_n) \ + \ 0(h^{q+2}) \ ,$$

onde q é a ordem do método. O termo $C_{q+1} h^{q+1} y^{(q+1)}(n_n)$ é frequentemente chamado de **Erro de Truncamento Local Principal**.

Assim, o erro de truncamento local, para :

a) o método de Euler é dado por:

$$\frac{h^2}{2!} \ y''(\xi) \ , \ \text{onde } x_n < \xi < x_{n+1} \ ,$$

isto é, o erro de truncamento local é da $O(h^2)$, e este é identicamente nulo se a solução de (12.2) é um polinômio de grau não excedendo 1.

b) o método do trapézio é dado por:

$$-\frac{h^3}{12} y'''(\xi) \ , \ \text{onde } x_n < \xi < x_{n+1} \ ,$$

isto é, o erro de truncamento local é da $O(h^3)$, o que representa um aperfeiçoamento sobre o método de Euler. Observe que o erro de truncamento local é exatamente o erro da regra do trapézio, fórmula (11.14), visto que o lado esquerdo da expressão (12.10) é calculada exatamente.

Exercício

12.4 - *Determine o erro de truncamento local para:*

a) *regra do ponto médio,*

b) *método de Simpson,*

c) *método de Adams-Moulton,*

d) *método de Adams-Bashforth,*

e) *método $\frac{3}{8}$ de Simpson.*

As propriedades mais importantes dos métodos numéricos para resolver problemas de valor inicial são consistência e estabilidade.

12.3.5 Consistência e Estabilidade

Descreveremos aqui as propriedades de consistência e estabilidade dos métodos de k-passos. Dado o método linear de passo múltiplo (12.7), definimos, inicialmente:

$$\rho(\xi) \;=\; \sum_{j=0}^{k} \alpha_j \; \xi^j \quad \text{e} \quad \tau(\xi) \;=\; \sum_{j=0}^{k} \beta_j \; \xi^j \; ,$$

como sendo o primeiro e segundo polinômio característico, respectivamente.

Definição 12.5 - *Um método linear de passo múltiplo é* **estável** *se nenhuma raiz de $\rho(\xi)$ tem módulo maior do que 1 e toda raiz com módulo 1 é simples.*

Exemplo 12.7 - *Verificar se o método de Simpson é estável.*

Solução: Temos que o método de Simpson é dado por (12.13), de onde deduzimos que:

$$\alpha_0 = -1 \; , \quad \alpha_1 = 0 \; , \quad \alpha_2 = 1 \; .$$

Assim:

$$\rho(\xi) \;=\; \xi^2 \;-\; 1 \quad \rightarrow \quad \xi \;=\; \pm 1$$

Logo as raízes têm módulo 1 e são simples. Portanto o método de Simpson é estável.

Definição 12.6 - *Um método linear de passo múltiplo é* **consistente** *se tem ordem $q \;\geq\; 1$.*

Assim, por (12.18), vemos que um método linear de passo múltiplo é consistente se e somente se

$$\sum_{j=0}^{k} \alpha_j \;=\; 0 \qquad \text{e} \qquad \sum_{j=0}^{k} \beta_j \;=\; \sum_{j=0}^{k} j\alpha_j \tag{12.20}$$

Exemplo 12.8 - *Verificar se o método de Adams-Basforth é consistente.*

Solução: Temos que o método de Adams-Basforth é dado por (12.15), de onde deduzimos que:

$$\begin{array}{ll} \alpha_0 \;=\; 0 \; , & \beta_0 \;=\; -\frac{1}{2} \; , \\ \alpha_1 \;=\; -1 \; , & \beta_1 \;=\; \frac{3}{2} \; , \\ \alpha_2 \;=\; 1 \; , & \beta_2 \;=\; 0 \; . \end{array}$$

Assim:

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 + \alpha_2 \;\Rightarrow\; C_0 = 0 - 1 + 1 = 0 \; , \\ C_1 &= \alpha_1 + 2\alpha_2 - (\beta_0 + \beta_1 + \beta_2) \;\Rightarrow\; C_1 = -\frac{1}{2} + \frac{3}{2} + 0 = 0 \; , \end{aligned}$$

Assim, o método de Adams-Basforth é consistente.

Definição 12.7 - *Se o erro de truncamento local de um método de k-passos é: $C_{q+1} h^{q+1} y^{(q+1)}(n_n)$, então dizemos que o método é* **consistente de ordem** *q.*

Pelo Exemplo 12.6 , vemos que o método de Euler é consistente de ordem 1 e que o método do trapézio é consistente de ordem 2.

Exercício

12.5 - *Determine a ordem de consistência dos seguintes métodos:*

a) *regra do ponto médio,*

b) *método de Simpson,*

c) *método de Adams-Moulton,*

d) *método* $\frac{3}{8}$ *de Simpson.*

12.3.6 Convergência

O resultado mais importante sobre métodos de passo múltiplo é saber se a aplicação de um determinado método será convergente para a solução exata do problema de valor inicial. Seja o **(p.v.i.)** (12.2), cuja solução exata é $y(x)$ e seja o método linear de passo múltiplo (12.7).

Por convergência entendemos que os valores encontrados convergem para a solução exata do problema, isto é, que $y_n \rightarrow y(x_n)$ quando $h \rightarrow 0$.

Definição 12.8 - *Um método linear de passo múltiplo é* **convergente** *se a seguinte afirmação é verdadeira: Seja $f(x, y)$ satisfazendo as condições do Teorema 12.1. Se $y(x)$ é solução do (p.v.i.) (12.2), então:*

$$\lim_{\substack{h \to 0 \\ hn = x - a (\text{fixo})}} y_n = y(x_n) \ ,$$

vale para todo $x \in [a, b]$ e todas as soluções y_n do método de passo múltiplo tendo valores iniciais y_μ satisfazendo $\lim_{h \to 0} y_\mu = y_0, \mu = 0, 1, \ldots, k-1$.

Assim para dar uma idéia de convergência, consideremos que estamos resolvendo um **(p.v.i.)**, com os seguintes comprimentos de passo: $h = h_0$, $\frac{1}{2}h_0$, $\frac{1}{4}h_0$, e $\bar{x} - a$ fixo, como mostrado na Figura 12.1.

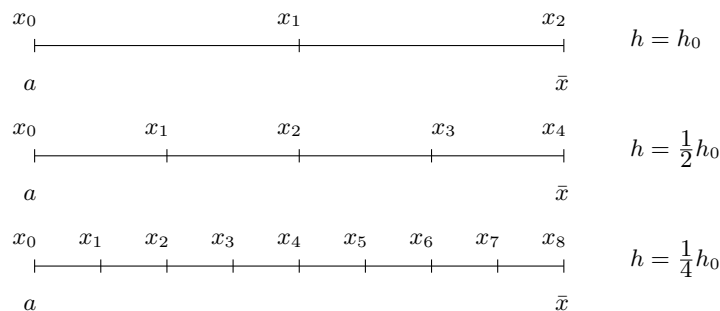



Figura 12.1

Seja $y_n(h)$ a notação para o valor de y_n obtido por um método numérico quando o tamanho do passo é h. Se estamos interessados, por exemplo, no valor de $y(x)$ quando $x = \bar{x}$, (Figura 12.1), teremos convergência se a sequência $y_2(h_0), y_4(\frac{1}{2}h_0), y_8(\frac{1}{4}h_0)$, convergir para o valor de $y(\bar{x})$, ou seja a verificação da convergência deve ser feita nos pontos da malha. Em geral, consideramos o caso em que h tende

continuamente a zero, isto é, consideramos $h = 0.1, 0.01, \ldots$.

Antes de definirmos as condições que garantem a convergência dos métodos de k-passos, analisemos o seguinte: quando calculamos o erro de truncamento local de um método de k-passos, intuitivamente, esperamos que tal erro ocorra pela aplicação do método linear de passo múltiplo num passo simples, ou seja que o erro ocorra apenas no cálculo de y_n, pois consideramos na análise do erro que a solução nos pontos anteriores são calculados exatamente. Entretanto, no cálculo de y_n, n passos (aproximadamente) são usados. Portanto se o erro de truncamento local for da $O(h^{q+1})$, o erro em y_n será:

$$nO(h^{q+1}) \;=\; nhO(h^q) \;=\; (x_n - x_0)O(h^q) \;.$$

Assim, se $h \to 0$ com x_n fixo, o **erro global** $y(x_n) - y_n$ é da $O(h^q)$.

Definição 12.9 - *Um método linear de passo mútiplo é convergente de ordem q, se o erro:*

$$y(x_n) - y_n \;=\; O(h^q) \;,$$

quando $h \to 0$, com x_n fixo.

Apresentamos assim uma idéia intuitiva de que se um método é consistente de ordem q então ele é convergente de ordem q. Entretanto, podemos enunciar o seguinte teorema, o qual pode ser provado rigorosamente.

Teorema 12.2 - *Um método linear de passo múltiplo é* **convergente** *de ordem q se e somente se é estável e consistente de ordem q.*

Prova: A prova deste teorema pode ser encontrada em [Henrici, 1962].

Assim, tanto a consistência como a estabilidade de um método de k-passos são importantes para garantir a convergência. Cabe salientar que enquanto a consistência controla o erro local em cada passo a estabilidade controla a forma pela qual o erro se propaga quando o número de passos aumenta. Além disso, quanto maior for a ordem de consistência do método, mais rapidamente obteremos a solução desejada.

Exemplo 12.9 - *Considere o* **(p.v.i.)***:*

$$\begin{cases} y' \;=\; y & ; \\ y(0) \;=\; 1 \qquad 0 \le x \le 1 \quad ; \quad h \;=\; 0.1 \end{cases}$$

cuja solução exata é $y(x) = e^x$. Verifique que, usando o seguinte método:

$$y_{n+2} \;=\; -3y_n + 4y_{n+1} - 2hf_n \;. \tag{12.21}$$

com $y_0 = 1$, e $y_1 = 1.10517$ não obtemos a solução do problema original. Analise então as condições que garantem a convergência.

Solução: A tabela a seguir mostra alguns dos valores obtidos com a aplicação do método:

x_n	y_n	$y(x_n)$
0	1	1
0.1	1.10517	1.10517
0.2	1.22068	1.22140
0.5	1.60638	1.64872
0.7	1.63634	2.01375
0.9	-0.74079	2.45960
1.0	-6.55860	2.71828

Vemos pelos valores obtidos que o método não é convergente. Analisemos então a consistência e a estabilidade. De (12.21), segue que:

$$y_{n+2} - 4y_{n+1} + 3y_n = -2hf_n .$$

e assim:

$$\alpha_0 = 3 , \qquad \beta_0 = -2 ,$$

$$\alpha_1 = -4 , \qquad \beta_1 = 0 ,$$

$$\alpha_2 = 1 , \qquad \beta_2 = 0 ,$$

Portanto:

$$C_0 = \alpha_0 + \alpha_1 + \alpha_2 \Rightarrow C_0 = 3 - 4 + 1 = 0 ,$$

$$C_1 = \alpha_1 + 2\alpha_2 - \beta_0 \Rightarrow C_1 = -4 + 2 + 2 = 0 .$$

Logo o método é consistente. Mas,

$$\rho(\xi) = \xi^2 - 4\xi + 3 = (\xi - 1)(\xi - 3) .$$

e assim as raízes de $\rho(\xi)$ são $\xi = 1$ e $\xi = 3$. Portanto o método não é estável. Isso explica a não convergência do método.

Exercícios

12.6 - *Verifique se o método explícito de dois passos:*

$$y_{n+2} - y_{n+1} = \frac{h}{3}[-2f_n + 3f_{n+1}] ,$$

pode ser utilizado para resolver um **(p.v.i.)** *com garantia de convergência.*

12.7 - *Mostre que o método implícito de dois passos:*

$$y_{n+2} - y_{n+1} = \frac{h}{12}[-f_n + 8f_{n+1} + 4f_{n+2}] ,$$

não é consistente.

12.4 Métodos do Tipo Previsor - Corretor

Descreveremos aqui como utilizar um método linear de passo múltiplo implícito, para determinar a solução do **(p.v.i.)** (12.2).

Para os métodos de k-passos implícitos, em cada passo, devemos resolver para y_{n+k} a equação:

$$y_{n+k} \; = \; - \sum_{j=0}^{k-1} \alpha_j y_{n+j} \; + \; h \sum_{j=0}^{k-1} \beta_j f_{n+j} \; + \; h \beta_k f(x_{n+k}, y_{n+k}) \; , \tag{12.22}$$

onde y_{n+j} e f_{n+j} , $j = 0, 1, \ldots, k-1$ são conhecidos.

Como já dissemos anteriormente, se f for uma função não linear em y, não teremos, em geral, condições de resolver (12.22) em relação a y_{n+k} de uma forma exata. Entretanto pode ser provado que uma única solução para y_{n+k} existe e pode ser aproximada pelo método iterativo:

$$y_{n+k}^{[s+1]} \; = \; - \sum_{j=0}^{k-1} \alpha_j y_{n+j} \; + \; h \sum_{j=0}^{k-1} \beta_j f_{n+j} \; + \; h \beta_k f(x_{n+k}, y_{n+k}^{[s]}) \; , \tag{12.23}$$

onde $s = 1, 2, \ldots,$ e mantendo x_{n+k} fixo, $y_{n+k}^{[0]}$, pode ser obtido usando um método linear de passo múltiplo explícito. Assim,

$$y_{n+k}^{[0]} \; = \; - \sum_{j=0}^{k-1} \alpha_j^* y_{n+j} \; + \; h \sum_{j=0}^{k-1} \beta_j^* f_{n+j} \; .$$

Ao método explícito chamaremos **Previsor**.

Com esse valor e o método implícito, (12.23), o qual chamaremos **Corretor**, calculamos $y_{n+k}^{[1]}$, $y_{n+k}^{[2]}, \ldots$.

Indicaremos por:

P**:** aplicação do Previsor,

E**:** cálculo de $f(x_{n+k}, y_{n+k}^{[s]})$,

C**:** aplicação do Corretor.

O par PC será então aplicado no modo $P(EC)^m E$, onde m é o número de vezes que calculamos f e aplicamos C. A iteração finaliza quando dois valores sucessivos de y, obtidos com a aplicação de C, satisfazem a precisão desejada.

Duas questões que surgem naturalmente vinculadas as fórmulas corretoras são:

1 - Sob que condições convergirá a fórmula corretora?

2 - Quantas iterações serão necessárias para se atingir a precisão desejada?

A resposta à última pergunta dependerá de muitos fatores. Contudo, a experiência mostra que somente uma ou duas aplicações da corretora são suficientes, desde que a amplitude do intervalo h tenha sido selecionada adequadamente. Caso verifiquemos que uma ou duas correções não são suficientes, será melhor reduzirmos a amplitude do intervalo h ao invés de prosseguirmos a iteração. Assim, na prática, não usamos $m > 2$.

A resposta à primeira questão está contida no seguinte teorema.

Teorema 12.3 - *Se $f(x,y)$ e $\frac{\partial f}{\partial y}$ forem contínuas em x e y no intervalo fechado $[a,b]$, e se $\frac{\partial f}{\partial y}$ não se anular neste intervalo, (12.23) convergirá, desde que h seja escolhido de modo a satisfazer:*

$$h < \frac{2}{\left|\frac{\partial f}{\partial y}\right|} .$$

Prova: Pode ser encontrada em [Conte,19..].

Podemos agora definir formalmente a aplicação do par PC, no modo $P(EC)^m E$: Calcular a cada passo:

$$y_{n+k}^{[0]} + \sum_{j=0}^{k-1} \alpha_j^* y_{n+j}^{[m]} = h \sum_{j=0}^{k-1} \beta_j^* f_{n+j}^{[m]} ,$$

para $s = 0, 1, \ldots, m-1$,

$$\begin{cases} f_{n+k}^{[s]} = f(x_{n+k}, y_{n+k}^{[s]}) \\ y_{n+k}^{[s+1]} = -\sum_{j=0}^{k-1} \alpha_j y_{n+j}^{[m]} + h\sum_{j=0}^{k-1} \beta_j f_{n+j}^{[m]} + h\beta_k f_{n+k}^{[s]} \end{cases} .$$

e finalmente,

$$f_{n+k}^{[m]} = f(x_{n+k}, y_{n+k}^{[m]}) .$$

Exemplo 12.10 - *Resolver o* **(p.v.i.)** *do exemplo 12.1, usando o par PC, onde:*

$$\begin{aligned} P: \quad y_{n+2} &= y_{n+1} + \frac{h}{2}[-f_n + 3f_{n+1}], \\ C: \quad y_{n+2} &= y_n + \frac{h}{3}[f_n + 4f_{n+1} + f_{n+2}] , \end{aligned} \tag{12.24}$$

no modo $P(EC)E$. Obter os valores iniciais necessários pelo método de Taylor de ordem 3.

Solução: Temos que: $y_0 = 2$ e, pelo exemplo 12.1, $y_1 = 2.0048$. Assim, fazendo $n = 0$ em (12.24), obtemos:

$$\begin{aligned} P: \ y_2^{(0)} &= y_1 + \frac{h}{2}[3f_1 - f_0] = 2.0048 + \frac{(0.1)}{2}[-0 + 3(0.0952)] \\ &= 2.0191 , \end{aligned}$$

desde que, pelo exemplo 12.1, $f_0 = 0$ e $f_1 = 0.0952$. Agora,

$$\begin{aligned} E: \ f_2^{(0)} &= f(x_2, y_2^{(0)}) = f(0.2, 2.0191) = -2.0191 + 0.2 + 2 \\ &= 0.1809 . \end{aligned}$$

Portanto:

$$\begin{aligned} C: \ y_2^{(1)} &= y_0 + \frac{h}{3}[f_0 + 4f_1 + f_2^{(0)}] = 2 + \frac{(0.1)}{3}[0 + 4(0.0952) + 0.1809] \\ &= 2.0187 \simeq y(x_2) = y(0.2) . \end{aligned}$$

Agora:

$$\begin{aligned} E: \ f_2^{(1)} &= f(x_2, y_2^{(1)}) = f(0.2, 2.0187) = -2.0187 + 0.2 + 2 \\ &= 0.1813 \ . \end{aligned}$$

Finalmente, fazendo $n = 1$ em (12.24), obtemos:

$$\begin{aligned} P: \ y_3^{(0)} &= y_2 \ + \ \frac{h}{2}[-f_1 + 3f_2] = 2.0187 + \frac{(0.1)}{2}[-0.0952 + 3(0.1813)] \\ &= 2.0411 \ , \end{aligned}$$

desde que $f_1 = 0.0952$ e $f_2^{(1)} = 0.1813$. Agora,

$$\begin{aligned} E: \ f_3^{(0)} &= f(x_3, y_3^{(0)}) = f(0.3, 2.0411) = -2.0411 + 0.3 + 2 \\ &= 0.2589 \ . \end{aligned}$$

Portanto:

$$\begin{aligned} C: \ y_3^{(1)} &= y_1 \ + \ \frac{h}{3}[f_1 + 4f_2 + f_3^{(0)}] \\ &= 2.0048 + \frac{(0.1)}{3}[0.0952 + 4(0.1809) + 0.2589] \\ &= 2.0407 \ \simeq y(x_3) \ = \ y(0.3) \ . \end{aligned}$$

Assim, a solução do **(p.v.i.)** dado é:

x_n	y_n
0.0	2.0000
0.1	2.0048
0.2	2.0187
0.3	2.0407

Compare os resultados obtidos com a solução obtida nos exemplos anteriores. Lembre-se que a solução exata do **(p.v.i.)** é: $y(x) = e^{-x} + x + 1$.

12.4.1 Erro de Truncamento Local

Supomos a aplicação do par PC, no modo $P(EC)^m$ ou $P(EC)^m E$ onde o previsor tem ordem $q^* \geq 0$; o corretor tem ordem $q \geq 1$ e $m \geq 1$. Pode-se mostrar que: (ver [Lambert, 19..])

1) se $q^* \geq q$ então o erro de truncamento local principal do par PC é o mesmo do C.

2) se $q^* = q - j;\ 0 < j \leq q$ então o erro de truncamento local principal do par PC é:

2.1) o mesmo do C se $m \geq j + 1$,

2.2) da mesma ordem do C, mas diferente dele se $m = j$,

2.3) da forma $kh^{q-j+m+1} + 0(h^{q-j+m+2})$ se $m \leq j - 1$.

Exercícios

12.8 - *Resolver o seguinte* **(p.v.i.)***:*

$$\begin{cases} y' = x^2 + y \\ y(0) = 1 ; \qquad x \in [0, 0.4];\ h = 0.2 \end{cases}$$

usando o par PC, onde:

$$P: \quad y_{n+1} = y_n + h\ f_n ,$$

$$C: \quad y_{n+1} = y_n + \frac{h}{2}[f_n + f_{n+1}] ,$$

no modo $P(EC)^2E$.

12.9 - *Considere o* **(p.v.i.)***:*

$$\begin{cases} y' = y(x - y) + 1 \\ y(0) = 1 ; \qquad x \in [0, 0.3];\ h = 0.15 \end{cases}$$

Resolva-o usando o par PC, onde:

$$P: \quad y_{n+2} = y_{n+1} + \frac{h}{2}[-f_n + 3f_{n+1}].$$

$$C: \quad y_{n+2} = y_{n+1} + \frac{h}{12}[-f_n + 8f_{n+1} + 5f_{n+2}]$$

no modo $P(EC)E$, e o método de Taylor de ordem 3 para obter os valores iniciais necessários.

12.10 - *Resolver o* **(p.v.i.)***:*

$$\begin{cases} y' = -2xy \\ y(0) = 1 ; \qquad x \in [0, 0.5];\ h = 0.1 \end{cases}$$

usando o par PC, onde:

$$P: \quad y_{n+3} = y_{n+2} + \frac{h}{12}[5f_n - 16f_{n+1} + 23f_{n+2}] ,$$

$$C: \quad y_{n+2} = y_n + \frac{h}{3}[f_n + 4f_{n+1} + f_{n+2}] ,$$

no modo $P(EC)^2E$, e o método de Taylor de ordem 3 para obter os valores iniciais necessários.

12.5 Método Geral Explícito de 1-passo

Muitas vezes desejamos resolver o **(p.v.i.)** (12.2) usando um método de k-passos; $k > 1$. Precisamos então obter os valores iniciais necessários, para se utilizar tal método, que sejam o mais preciso possível. Isto pode ser feito através do método de Taylor de ordem q, se possível, pois nem sempre existem as derivadas de ordem superior da f, ou então pelos métodos de Runge-Kutta, desde que ambos são métodos explícitos de 1-passo. Os métodos de Runge-Kutta, assim como o método de Taylor de ordem q, também podem ser utilizados para determinar a solução do **(p.v.i.)** para $x \in [a, b]$.

Definição 12.10 - *Um* **método geral explícito de 1-passo** *é definido pela relação:*

$$y_{n+1} \; - \; y_n \; = \; h\phi(x_n, y_n, h) \; . \tag{12.25}$$

onde ϕ é uma função que depende de x_n, y_n e h.

12.5.1 Ordem

Definição 12.11 - *O método (12.25) é de* **ordem** *q, se q é o maior inteiro tal que:*

$$y(x+h) \; - \; y(x) \; - \; h\phi(x, y(x), h) \; = \; 0(h^{q+1}) \; , \tag{12.26}$$

onde $y(x)$ é a solução exata do **(p.v.i.)** *(12.2).*

12.5.2 Consistência

Definição 12.12 - *O método (12.25) é* **consistente** *com o* **(p.v.i.)** *(12.2) se:*

$$\phi(x, y, 0) \; = \; f(x, y) \; . \tag{12.27}$$

Exemplo 12.11 - *Considere o método de Taylor de ordem q, dado por (12.5).*

a) *Verificar que (12.5) é um método geral explícito de um passo.*

b) *Determinar sua ordem, usando (12.26).*

c) *Verificar se é consistente, usando (12.27).*

Solução: Temos por (12.5), que:

$$\begin{aligned} y_{n+1} \quad &= \quad y_n + h\, f_n + \frac{h^2}{2!}\, f'_n + \ldots + \frac{h^q}{q!}\, f_n^{(q-1)} \\ &= \quad y_n + h \left[f_n + \frac{h}{2!}\, f'_n + \ldots + \frac{h^{q-1}}{q!}\, f_n^{(q-1)} \right] \\ &= \quad y_n + h\phi_T(x_n, y_n, h) \; , \end{aligned}$$

onde denotamos por $\phi_T(x, y, h)$, a função ϕ do método de Taylor calculada no ponto (x, y), isto é:

$$\phi_T(x, y, h) \; = \; f(x, y) + \frac{h}{2!}\, f'(x, y) + \ldots + \frac{h^{q-1}}{q!}\, f^{(q-1)}(x, y) \; . \tag{12.28}$$

Assim (12.5) é um método geral explícito de um passo. Agora,

$$\begin{aligned} & y(x+h) - \ y(x) - h\phi(x,y,h) \\ = \ & y(x) + hy'(x) + \frac{h}{2!}y''(x) + \ldots + \frac{h^q}{q!}y^{(q)}(x) + O(h^{q+1}) \\ - \ & y(x) - h\left[f(x,y) + \frac{h}{2!}f'(x,y) + \ldots + \frac{h^{q-1}}{q!}\ f^{(q-1)}(x,y) + O(h^q)\right] \\ = \ & O(h^{q+1})\ . \end{aligned}$$

onde desenvolvemos $y(x+h)$ em série de Taylor em torno do ponto x e substituimos $\phi(x,y,h)$ pela ϕ_T . Portanto o método de Taylor tem ordem q e é consistente com o **(p.v.i)** (12.2), pois $\phi(x,y,0) = f(x,y)$.

12.5.3 Convergência

Teorema 12.4 - *Seja $\phi(x,y,h)$ satisfazendo as condições:*

i) $\phi(x,y,h)$ é contínua em

$$S = \{(x,y,h)\ a \leq x \leq b\ ;\ -\infty\ < y < \infty\ ;\ 0 < h \leq h_0, h_0 > 0\}\ .$$

ii) $\phi(x,y,h)$ satisfaz a condição de Lipschitz em relação a y, isto é:

$$|\phi(x,y,h)\ -\ \phi(x,y^*,h)|\ \leq\ L|y-y^*|\ ,$$

para todos os pontos (x,y,h) e (x,y^*,h) em S.

Então o método (12.25) é **convergente** *se e somente se é consistente.*

Prova: A prova deste teorema pode ser encontrada em ...

Para todos os métodos que estudaremos aqui, as condições **i)** e **ii)** do Teorema 12.4 são satisfeitas se $f(x,y)$ satisfaz as hipóteses do Teorema 12.1. Para tais métodos consistência é condição necessária e suficiente para garantir convergência. Note que não existe exigência correspondente a estabilidade, desde que nenhuma solução *parasítica* pode ocorrer com método de um passo.

Exercícios

12.11 - *Considere o método de Euler, dado por (12.8).*

a) *Verificar se (12.8) é um método geral explícito de um passo.*

b) *Determinar sua ordem, usando (12.26).*

c) *Verificar se é consistente, usando (12.27).*

12.12 - *Mostre que um método do tipo (12.25) consistente tem ordem pelo menos 1.*

Estamos agora em condições de definir os métodos de Runge-Kutta.

12.5.4 Métodos de Runge-Kutta

Definição 12.13 - *O* **método geral de Runge-Kutta** *de* **R** *estágios é definido por:*

$$y_{n+1} \ - \ y_n \ = \ h\phi(x_n, y_n, h) \ ,$$

onde

$$\phi(x,y,h) \ = \ \sum_{r=1}^{R} c_r k_r \ , \tag{12.29}$$

$$k_1 \ = \ f(x,y) \ ,$$

$$k_r \ = \ f(x + a_r h \ , \ y + h \sum_{s=1}^{r-1} b_{rs} k_s) \ ; \ r = 2,3,\ldots,R \ ,$$

$$a_r \ = \ \sum_{s=1}^{r-1} b_{rs} \ ; \ r = 2,3,\ldots,R \ .$$

Para se obter métodos de Runge-Kutta devemos determinar as constantes c_r, a_r e b_{rs} da definição 12.13. Determinamos estas constantes comparando a expansão da função $\phi(x,y,h)$, definida por (12.29), em potências de h, com a função $\phi_T(x,y,h)$ do método de Taylor, (12.28), no sentido de se obter métodos de determinada ordem. Veremos a seguir como fazer isso.

Métodos de Runge-Kutta de ordem 2

Consideremos incialmente, que desejamos obter métodos de Runge-Kutta de 2 estágios. Devemos tomar então $R = 2$, na definição (12.13). Assim:

$$\phi(x,y,h) \ = \ c_1 k_1 \ + \ c_2 k_2 \ ,$$

$$k_1 \ = \ f(x,y) \ ,$$

$$k_2 \ = \ f(x + a_2 h \ , \ y + h b_{21} k_1) \ ,$$

$$a_2 \ = \ b_{21} \ .$$

Portanto:

$$k_2 \ = \ f(x + a_2 h \ , \ y \ + \ h a_2 f) \ .$$

Desenvolvendo k_2 em série de Taylor em torno do ponto (x,y), obtemos:

$$k_2 \ = \ f(x,y) + (a_2 h) f_x(x,y) + (h a_2 f) f_y(x,y) + \frac{(a_2 h)^2}{2!} f_{xx}(x,y)$$

$$+ \ (a_2 h)(h a_2 f) f_{xy}(x,y) + \frac{(h a_2 f)^2}{2!} f_{yy}(x,y) + O(h^3) \ .$$

Portanto:

$$\begin{aligned}
\phi(x,y,h) &= c_1k_1 + c_2k_2 \\
&= c_1 f + c_2 \left[f + (a_2h)f_x + (a_2hf)f_y + \frac{(a_2h)^2}{2!} f_{xx} \right. \\
&+ \left. (a_2h)^2 f f_{xy} + \frac{(a_2hf)^2}{2!} f_{yy} + O(h^3) \right] \\
&= (c_1 + c_2) f + c_2 a_2 h (f_x + f_y f) \\
&+ \frac{(a_2h)^2}{2!} c_2 \left[f_{xx} + 2 f f_{xy} + f_{yy} f^2 \right] + O(h^3) \ ,
\end{aligned}$$

onde agrupamos os termos de mesma potência de h. Denotando por:

$$F \ = \ f_x + f_y f \ \text{ e } \ G \ = \ f_{xx} + 2 f f_{xy} + f_{yy} f^2 \ . \tag{12.30}$$

obtemos:

$$\phi(x,y,h) = (c_1 + c_2) f + c_2 a_2 h F + \frac{(a_2h)^2}{2!} c_2 G + 0(h^3) \ , \tag{12.31}$$

Agora, podemos escrever a função $\phi_T(x,y,h)$, (12.28), como:

$$\begin{aligned}
\phi_T(x,y,h) &= f(x,y) + \frac{h}{2!} f'(x,y) + \frac{h^2}{3!} f''(x,y) + O(h^3) \\
&= f + \frac{h}{2!}(f_x + f_y f) + \frac{h^2}{3!}(f_{xx} + 2 f_{xy} \ f \\
&+ f_{yy} f^2 + f_x f_y + f_y^2 \ f) + O(h^3)
\end{aligned}$$

Agrupando os termos semelhantes e usando (12.30), obtemos:

$$\phi_T(x,y,h) = f \ + \ \frac{h}{2} F \ + \frac{h^2}{3!} \left[G + f_y F \right] \ + \ 0(h^3) \tag{12.32}$$

Para determinar um método de 2 estágios e ordem máxima, comparamos (12.31) com (12.32), obtendo:

$$\begin{cases} c_1 \ + \ c_2 \ = \ 1 \\ c_2 a_2 \ = \ \frac{1}{2} \end{cases} \tag{12.33}$$

Resolvendo esse sistema iremos obter métodos de Runge-Kutta de ordem 2, pois na definição (12.13), temos $h\phi(x,y,h)$ e portanto estamos impondo igualdade até termos da $O(h^2)$. Além disso, como o sistema (12.33) possui 2 equações e 3 incógnitas, este possui infinitas soluções e portanto podemos afirmar que existem infinitos métodos de Runge-Kutta de 2 estágios e ordem 2.

Observe que para se obter um método de Runge-Kutta de 2 estágios e ordem 3, é necessário que além de (12.33) tenhamos:

$$\begin{aligned} \frac{a_2^2 c_2}{2} G &= \frac{1}{6} + \frac{1}{6} f_y F \\ \Rightarrow \quad \left(\frac{a_2^2 c_2}{2} G - \frac{1}{6} \right) &= \frac{1}{6} f_y F \ . \end{aligned}$$

A igualdade acima só pode ser satisfeita impondo severas condições sobre a função f, e portanto não existem métodos de Runge-Kutta de 2 estágios e ordem 3.

Assim, atribuindo um valor para uma das constantes em (12.33), obtemos as outras duas, em função desta. Os métodos de Runge-Kutta de 2-estágios e ordem 2, mais usados são obtidos tomando:

a) $c_1 = 0 \Rightarrow c_2 = 1$ e $a_2 = \frac{1}{2}$. Portanto:

$$\begin{aligned} y_{n+1} &= y_n + h k_2 \text{ , onde :} \\ k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + \tfrac{1}{2} h, y_n + \tfrac{1}{2} h k_1) \ , \end{aligned} \tag{12.34}$$

que é conhecido como **Método de Euler Modificado**. Observe que apesar de k_1 não aparecer explicitamente, ele deve ser calculado a cada passo.

b) $c_1 = \frac{1}{2} \Rightarrow c_2 = \frac{1}{2}$ *e* $a_2 = 1$. Portanto:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \text{ , onde :} \\ k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + h, y_n + h k_1) \ , \end{aligned} \tag{12.35}$$

que é conhecido como **Método de Euler Melhorado**.

Exemplo 12.12 *-Usando o método de Euler Modificado, resolva o* **(p.v.i.)** *do exemplo 12.1.*

Solução: Temos que $y_0 = 2$, (condição inicial). Fazendo $n = 0$, em (12.34), obtemos:

$$y_1 = y_0 + h k_2 \ ,$$

onde:

$$\begin{aligned} k_1 &= f(x_0, y_0) = f(0, 2) = -2 + 0 + 2 = 0 \ , \\ k_2 &= f\left(x_0 + \frac{1}{2} h, y_0 + \frac{1}{2} h k_1 \right) = f\left(0 + \frac{0.1}{2}, 2 + \frac{0.1}{2}(0) \right) \\ &= f(0.05, 2) = -2 + 0.05 + 2 = 0.05 \ . \end{aligned}$$

Portanto:
$$y_1 \;=\; 2 + 0.1(0.05) \;=\; 2.005 \simeq y(x_1) = y(0.1)\ .$$
Fazendo agora $n = 1$, em (12.34), obtemos:
$$y_2 \;=\; y_1 + hk_2\ ,$$
onde:
$$\begin{aligned} k_1 &= f(x_1, y_1) = f(0.1, 2.005) \ = -2.005 + 0.1 + 2 = 0.095\ , \\ k_2 &= f\left(x_1 + \frac{1}{2}h, y_1 + \frac{1}{2}hk_1\right) = f\left(0.1 + \frac{0.1}{2}, 2.005 + \frac{0.1}{2}(0.095)\right) \\ &= f(0.15, 2.0098) = -2.0098 + 0.15 + 2 = 0.1403\ . \end{aligned}$$
Portanto:
$$y_2 \;=\; 2.005 + 0.1(0.1403) \;=\; 2.0190 \simeq y(x_2) = y(0.2)\ .$$
Finalmente, fazendo $n = 2$, em (12.34), obtemos:
$$y_3 \;=\; y_2 + hk_2\ ,$$
onde:
$$\begin{aligned} k_1 &= f(x_2, y_2) = f(0.2, 2.0190) \ = -2.0190 + 0.2 + 2 = 0.1810\ , \\ k_2 &= f\left(x_2 + \frac{1}{2}h, y_2 + \frac{1}{2}hk_1\right) = f\left(0.2 + \frac{0.1}{2}, 2.0190 + \frac{0.1}{2}(0.1810)\right) \\ &= f(0.25, 2.0281) = -2.0281 + 0.25 + 2 = 0.2220\ . \end{aligned}$$
Portanto:
$$y_3 \;=\; 2.0190 + 0.1(0.2220) \;=\; 2.0412 \simeq y(x_3) = y(0.3)\ .$$
Assim, a solução do **(p.v.i.)** dado é:

x_n	y_n
0.0	2.00000
0.1	2.005
0.2	2.0190
0.3	2.0412

Métodos de Runge-Kutta de ordem 3

Se desejamos obter métodos de Runge-Kutta de 3 estágios, devemos além do que já foi feito na seção anterior, desenvolver também k_3 em série de Taylor, pois os métodos de Runge-Kutta de 3 estágios são obtidos a partir de:
$$y_{n+1} \;=\; y_n + h(c_1k_1 + c_2k_2 + c_3k_3)\ ,$$

onde, k_1 e k_2 possuem as mesmas expressões do método de 2-estágios e,

$$\begin{aligned} k_3 &= f(x + ha_3, y + hb_{31}k_1 + b_{32}k_2) \\ &= f(x + ha_3, y + h(a_3 - b_{32})k_1 + b_{32}k_2) \ , \end{aligned}$$

desde que $a_3 = b_{31} + b_{32}$. Devemos então agrupar os termos semelhantes e compará-los com a $\phi_T(x, y, h)$. Como pode ser observado na seção anterior, a obtenção de métodos de Runge-Kutta envolve manipulações tediosas, e assim serão omitidas. Daremos aqui apenas o sistema obtido quando se compara ϕ com ϕ_T para se obter métodos de Runge-Kutta de 3 estágios e ordem máxima. Assim:

$$\left\{ \begin{array}{l} c_1 + c_2 + c_3 = 1 \\ c_2 a_2 + c_3 a_3 = \frac{1}{2} \\ c_3 b_{32} a_2 = \frac{1}{6} \\ c_2 a_2^2 + c_3 a_3^2 = \frac{1}{3} \end{array} \right. \tag{12.36}$$

que é um sistema de 4 equações e 6 incógnitas, onde comparamos os termos de ϕ e ϕ_T até $O(h^3)$. Atribuindo valores a duas das variáveis obtemos as outras quatro em função destas. Novamente temos infinitos métodos de Runge-Kutta de 3 estágios e ordem 3. Também nesse caso não conseguimos métodos de 3 estágios e ordem 4 a menos que se imponha condições sobre a f.

Os métodos de Runge-Kutta de 3 estágios e ordem 3, mais populares, são obtidos de (12.36), fazendo:

a) $c_1 = \frac{1}{4}$ e $c_2 = 0$.

Assim, da primeira equação, de (12.36), obtemos: $c_3 = \frac{3}{4}$. Substituindo na segunda equação segue que $\frac{3}{4}a_3 = \frac{1}{2} \rightarrow a_3 = \frac{2}{3}$. Finalmente da última equação, resulta que:

$$(0)(a_2^2) + \frac{3}{4}\left(\frac{2}{3}\right)^2 = \frac{1}{3} \ ,$$

que é satisfeita para qualquer valor de a_2. Escolhendo então $a_2 = \frac{1}{3}$ obtemos da terceira equação que $b_{32} = \frac{2}{3}$. Portanto:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}(k_1 + 3k_3) \ , \text{onde :} \\ k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1) \ , \\ k_3 &= f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2) \ , \end{aligned} \tag{12.37}$$

que é conhecido como **Método de Heun**. Novamente, o termo k_2 não aparece explicitamente, mas deve ser calculado a cada passo.

b) $c_2 \; = \; c_3$ e $a_2 \; = \; a_3$.

Substituindo os valores na segunda e quarta equações, segue que:

$$\begin{cases} 2c_3a_3 = \frac{1}{2} \Rightarrow c_3a_3 = \frac{1}{4} \\ \\ 2c_3a_3^2 = \frac{1}{3} \Rightarrow c_3a_3^2 = \frac{1}{6} \end{cases} \tag{12.38}$$

Substituindo em (12.38), a primeira equação na segunda obtemos que: $a_3 = \frac{2}{3} = a_2$. Assim $c_3 = \frac{3}{8} = c_2$. Da primeira equação obtemos: $c_1 = 1 - 2c_3 = 1 - \frac{3}{4} \rightarrow c_1 = \frac{1}{4}$. Finalmente, de $c_3b_{32}a_2 = \frac{1}{6}$ segue que: $b_{32} = \frac{2}{3}$. Portanto:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{4}\left[k_1 + \frac{3}{2}(k_2 + k_3)\right] \text{ , onde :} \\ k_1 &= f(x_n, y_n) \text{ ,} \\ k_2 &= f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_1) \text{ ,} \\ k_3 &= f(x_n + \frac{2}{3}h, y_n + \frac{2}{3}hk_2) \text{ ,} \end{aligned} \tag{12.39}$$

que é conhecido como **Método de Nystrom**.

Exemplo 12.13 - *Resolver o* **(p.v.i.)** *do exemplo 12.1, usando o par PC dado por 12.24, no modo $P(EC)$. Obtenha os valores iniciais necessários pelo método de Heun, fórmula (12.37).*

Solução: Temos que $y_0 = 2$, (condição inicial). Fazendo $n = 0$ em (12.37), obtemos:

$$y_1 = y_0 + \frac{h}{4}(k_1 + 3k_3) \text{ , onde :}$$

$$\begin{aligned} k_1 &= f(x_0, y_0) = f(0,2) = -2 + 0 + 2 = 0, \\ k_2 &= f(x_0 + \frac{1}{3}h, y_0 + \frac{1}{3}hk_1) = f(0 + \frac{0.1}{3}, 2 + \frac{0.1}{3}(0)), \\ &= f(0.0333, 2) = -2 + 0.0333 + 2 = 0.0333 \\ k_3 &= f(x_0 + \frac{2}{3}h, y_0 + \frac{2}{3}hk_2) = f(0 + \frac{0.2}{3}, 2 + \frac{0.2}{3}(0.0333)) \\ &= f(0.0667, 2.0022) = -2.0022 + 0.0667 + 2 = 0.0645 \end{aligned}$$

Assim

$$y_1 = 2 + \frac{0.1}{4}(0 + 3(0.0645)) = 2.0048 \simeq y(x_1) = y(0.1)$$

Desde que $y_1 = 2.0048$, a determinação de y_2 e y_3, usando o par PC dado por 12.24, no modo $P(EC)$, fornece exatamente o mesmo resultado do exemplo 12.10, ou seja , a solução do **(p.v.i.)** dado é:

x_n	y_n
0.0	2.0000
0.1	2.0048
0.2	2.0187
0.3	2.0407

Métodos de Runge-Kutta de ordem 4

Neste caso, a comparação de ϕ com ϕ_T, para se obter métodos de Runge-Kutta de 4 estágios e ordem máxima, fornece um sistema de 11 equações e 13 incógnitas. Cada solução desse sistema define um método de Runge-Kutta com ordem 4. Portanto existem infinitos métodos de Runge-Kutta de 4 estágios e ordem 4.

O dois métodos mais utilizados de Runge-Kutta de 4 estágios e ordem 4 são dados por:

$$\begin{aligned} y_{n+1} - y_n &= \frac{h}{6}[k_1 + 2(k_2+k_3) + k_4] \text{ , onde :} \\ k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_1) \ , \\ k_3 &= f(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hk_2) \ , \\ k_4 &= f(x_n + h, y_n + hk_3) \ , \end{aligned} \tag{12.40}$$

e

$$\begin{aligned} y_{n+1} - y_n &= \frac{h}{8}[k_1 + 3(k_2+k_3) + k_4] \text{ , onde :} \\ k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + \frac{1}{3}h, y_n + \frac{1}{3}hk_1) \ , \\ k_3 &= f(x_n + \frac{2}{3}h, y_n - \frac{1}{3}hk_1) + hk_2) \ , \\ k_4 &= f(x_n + h, y_n + hk_1 - hk_2 + hk_3) \ , \end{aligned} \tag{12.41}$$

Exemplo 12.14 - *Resolver o* **(p.v.i.)** *do exemplo 12.1, usando o método dado por (12.40).*

Solução: Temos que $y_0 = 2$. Fazendo $n = 0$ em (12.40), obtemos:

$$y_1 = y_0 + \frac{h}{6}[k_1 + 2(k_2+k_3) + k_4] \ ,$$

onde:

$$\begin{aligned}
k_1 &= f(x_0, y_0) = 0 \ , \\
k_2 &= f(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}hk_1) \ , \\
&= f(0 + \frac{0.1}{2}, 2 + \frac{0.1}{2}(0)) = f(0.05, 2) = 0.05 \ , \\
k_3 &= f(x_0 + \frac{1}{2}h, y_0 + \frac{1}{2}hk_2) \ , \\
&= f(0 + \frac{0.1}{2}, 2 + \frac{0.1}{2}(0.05)) = f(0.05, 2.0025) = 0.0475 \ , \\
k_4 &= f(x_0 + h, y_0 + hk_3) \ , \\
&= f(0 + 0.1, 2 + 0.1(0.0475)) = f(0.1, 2.0048) = 0.0952
\end{aligned}$$

Portanto:

$$\begin{aligned}
y_1 &= 2 + \frac{0.1}{6}[0 + \ 2(0.05 + 0.0475) + 0.0952] \\
&= 2.00484 \simeq y(x_1) = y(0.1) \ ,
\end{aligned}$$

Fazendo $n = 1$ em (12.40), obtemos:

$$y_2 = y_1 + \frac{h}{6}[k_1 + \ 2(k_2 + k_3) + k_4] \ ,$$

onde:

$$\begin{aligned}
k_1 &= f(x_1, y_1) = f(0.1, 2.00484) = 0.0952 \ , \\
k_2 &= f(x_1 + \frac{1}{2}h, y_1 + \frac{1}{2}hk_1) \ , \\
&= f(0.1 + \frac{0.1}{2}, 2.00484 + \frac{0.1}{2}(0.0952)) = f(0.15, 2.0096) = 0.1404 \ , \\
k_3 &= f(x_1 + \frac{1}{2}h, y_1 + \frac{1}{2}hk_2) \ , \\
&= f(0.1 + \frac{0.1}{2}, 2.00484 + \frac{0.1}{2}(0.1404)) = f(0.15, 2.0119) = 0.1381 \ , \\
k_4 &= f(x_1 + h, y_1 + hk_3) \ , \\
&= f(0.1 + 0.1, 2.00484 + 0.1(0.1381)) = f(0.2, 2.0187) = 0.1813 \ .
\end{aligned}$$

Portanto:

$$y_2 = 2.00484 + \frac{0.1}{6}[0.0952 + \ 2(0.1404 + 0.1381) + 0.1813]$$

$$= \quad 2.01873 \simeq y(x_2) = y(0.2) \ ,$$

Finalmente, fazendo $n = 2$ em (12.40), obtemos:

$$y_3 = y_2 + \frac{h}{6}[k_1 + \ 2(k_2 + k_3) + k_4] \ ,$$

onde:

$$\begin{aligned}
k_1 &= f(x_2, y_2) = f(0.2, 2.01873) = 0.1813 \ , \\
k_2 &= f(x_2 + \frac{1}{2}h, y_2 + \frac{1}{2}hk_1) \ , \\
&= f(0.2 + \frac{0.1}{2}, 2.01873 + \frac{0.1}{2}(0.1813)) = f(0.25, 2.0278) = 0.2222 \ , \\
k_3 &= f(x_2 + \frac{1}{2}h, y_2 + \frac{1}{2}hk_2) \ , \\
&= f(0.2 + \frac{0.1}{2}, 2.01873 + \frac{0.1}{2}(0.2222)) = f(0.25, 2.0298) = 0.2202 \ , \\
k_4 &= f(x_2 + h, y_2 + hk_3) \ , \\
&= f(0.2 + 0.1, 2.01873 + 0.1(0.2202)) = f(0.3, 2.0408) = 0.2592 \ .
\end{aligned}$$

Portanto:

$$\begin{aligned}
y_2 &= 2.01873 + \frac{0.1}{6}[0.1813 + \ 2(0.2222 + 0.2202) + 0.2592] \\
&= 2.04082 \simeq y(x_3) = y(0.3) \ ,
\end{aligned}$$

Assim a solução do **(p.v.i.)** dado é:

x_n	y_n
0	2
0.1	2.00484
0.2	2.01873
0.3	2.04082

Pelo que foi visto nessa seção nos dá a impressão que podemos obter sempre métodos de Runge-Kutta de R estágios e ordem R. Entretanto, [Butcher, 1964], provou a não existência de métodos de Runge-Kutta de 5 estágios e ordem 5. Além disso, provou o seguinte resultado: Seja $q(R)$ a maior ordem que

pode ser obtida por um método de Runge-Kutta de R estágios. Então:

$$\begin{aligned} q(R) &= R\ ,\ R = 1,2,3,4, \\ q(5) &= 4 \\ q(6) &= 5 \\ q(7) &= 6 \\ q(8) &= 6 \\ q(9) &= 7 \\ q(R) &\leq R-2\ ,\ R = 10,11,\ldots \end{aligned}$$

Na prática os métodos de Runge-Kutta mais utilizados são os de ordem 4.

Exercícios

12.13 - *Mostre que o método de Euler melhorado é equivalente a aplicação do método previsor-corretor, onde o previsor é o metodo de Euler e o corretor o método do trapézio, aplicados no modo $P(EC)E$.*

12.14 - *Resolva o* **(p.v.i.)** *do exemplo 12.1, usando o método da Adams-Basforth. Escolha um método de Runge-Kutta de ordem 2, para obter os valores iniciais necessários.*

12.15 - *Verifique que, usando o método de Euler modificado para resolver o* **(p.v.i.)***:*

$$\begin{cases} y' = y - x \\ \\ y(0) = 2\ ; \qquad x \in [0, 0.5];\ h = 0.1 \end{cases}$$

obtem-se:

x_n	y_n		$y(x) = e^x + x + 1$
	h = 0.1	*h = 0.01*	
0.0	*2.00000*	*2.00000*	*2.00000*
0.5	*3.14745*	*3.14870*	*3.14872*
1.0	*4.71408*	*4.71824*	*4.71828*

12.16 - *Resolva o* **(p.v.i.)** *do exemplo 12.1, usando o método de Nystrom.*

12.17 - *Obter um método de Runge-Kutta de 3 estágios e ordem 3, tomando em (12.36),* $c_1 = \dfrac{1}{6}$ e $c_2 = \dfrac{2}{3}$.

12.6 Sistemas de Equações e Equações de Ordem Elevada

Até agora nos preocupamos em resolver equações diferenciais de primeira ordem, mais especificamente problemas de valor inicial de primeira ordem. Entretanto, a maioria das equações diferenciais com importância prática, são de ordem maior que 1 ou então são sistemas de equações diferenciais. Veremos inicialmente como resolver um sistema de equações diferenciais de primeira ordem, e, para finalizar esse capítulo, como resolver numericamente uma equação diferencial de ordem elevada.

12.6.1 Sistemas de Equações Diferenciais

Consideremos um sistema de n equações diferenciais de primeira ordem:

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2, \ldots, y_n) \\ y_2' &= f_2(x, y_1, y_2, \ldots, y_n) \\ &\vdots \\ y_n' &= f_n(x, y_1, y_2, \ldots, y_n) \end{aligned}$$

o qual pode ser escrito, como:

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}) ,$$

onde $\mathbf{y}$, $\mathbf{y}'$ e $\mathbf{f}$ são vetores com componentes y_i, y_i' e $f_i, (i = 1, 2, \ldots, n)$, respectivamente. Para que esse sistema possua uma única solução é necessário impormos uma condição adicional sobre $\mathbf{y}$. Esta condição é usualmente da forma:

$$\mathbf{y}(x_0) = \mathbf{y}_0 ,$$

para um dado número x_0 e um vetor $\mathbf{y}_0$. Condições suficientes para a existência e unicidade de solução de tais sistemas podem ser encontradas em [Rao,1981].

Agora descreveremos como os métodos apresentados nas seções anteriores para a solução de equações diferenciais de primeira ordem podem ser aplicados para resolver sistemas de equações diferenciais de primeira ordem. Para efeito de simplicidade, e sem perda de generalidade, consideramos apenas o caso em que $n = 2$, isto é, o sistema possui apenas duas equações, e, para maior clareza usaremos a notação:

$$\begin{cases} y' = f(x, y, z) \\ z' = g(x, y, z) \\ y(x_0) = y_0 \\ z(x_0) = z_0 \qquad x \in [x_0, b] \end{cases} \tag{12.42}$$

Assim, se desejarmos resolver o sistema (12.42) pelo método de Euler, teremos:

$$\begin{aligned} y_{n+1} &= y_n + hf(x_n, y_n, z_n) \\ z_{n+1} &= z_n + hg(x_n, y_n, z_n) \end{aligned} \tag{12.43}$$

que será aplicado passo a passo, como mostra o exemplo a seguir.

Exemplo 12.15 - *Usando o método de Euler, resolver o seguinte sistema diferencial:*

$$\begin{cases} y' = z \\ z' = y + e^x \\ y(0) = 1 \\ z(0) = 0 \qquad x \in [0, 0.2] , \quad h = 0.1 \end{cases}$$

Solução: Para esse sistema, usando a fórmula (12.43), obtemos:

$$\begin{aligned} y_{n+1} &= y_n + h(z_n) \\ z_{n+1} &= z_n + h(y_n + e^{x_n}) \end{aligned}$$

Assim, fazendo $n = 0$, obtemos:

$$\begin{aligned} y_1 &= y_0 + h(z_0) = 1 + 0.1(0) \ = \ 1 \simeq y(x_1) = y(0.1) \\ z_1 &= z_0 + h(y_0 + e^{x_0}) = 0 + 0.1(1 + e^0) = 0.2 \simeq z(x_1) = z(0.1) \end{aligned}$$

Para $n = 1$, segue que:

$$\begin{aligned} y_2 &= y_1 + h(z_1) = 1 + 0.1(0.2) \ = \ 1.02 \simeq y(x_2) = y(0.2) \\ z_2 &= z_1 + h(-y_1) = 0.2 + 0.1(1 + e^{0.1}) = 0.4105 \simeq z(x_2) = z(0.2) \end{aligned}$$

Assim, a solução do sistema dado é:

x_n	y_n	z_n
0.0	1	0
0.1	1	0.2
0.2	1.02	0.4105

Exemplo 12.16 - *Resolver o sistema dado no exemplo 12.15, usando o par PC, onde:*

$$P\text{:} \quad y_{n+1} \ = \ y_n \ + \ h \ f_n \ ,$$

$$C\text{:} \quad y_{n+1} \ = \ y_n \ + \ \frac{h}{2}[f_n + f_{n+1}] \ ,$$

no modo $P(EC)E$.

Solução: Temos então:

$$P: \quad \begin{aligned} y_{n+1}^{(0)} &= y_n + hf_n \\ z_{n+1}^{(0)} &= z_n + hg_n \end{aligned}$$

$$E: \quad \begin{aligned} f_{n+1}^{(0)} &= f(x_{n+1}, y_{n+1}^{(0)}, z_{n+1}^{(0)}) \\ g_{n+1}^{(0)} &= g(x_{n+1}, y_{n+1}^{(0)}, z_{n+1}^{(0)}) \end{aligned}$$

$$C: \quad \begin{aligned} y_{n+1}^{(1)} &= y_n + \frac{h}{2}[f_n + f_{n+1}^{(0)}] \\ z_{n+1}^{(1)} &= z_n + \frac{h}{2}[g_n + g_{n+1}^{(0)}] \end{aligned}$$

$$E: \quad \begin{aligned} f_{n+1}^{(1)} &= f(x_{n+1}, y_{n+1}^{(1)}, z_{n+1}^{(1)}) \\ g_{n+1}^{(1)} &= g(x_{n+1}, y_{n+1}^{(1)}, z_{n+1}^{(1)}) \end{aligned}$$

No nosso problema:

$$\begin{aligned} f_n &= f(x_n, y_n, z_n) = z_n \ , \\ g_n &= g(x_n, y_n, z_n) = y_n + e^{x_n} \ , \\ f_{n+1}^{(k)} &= z_{n+1}^{(k)} \ , \quad k = 0, 1 \ , \\ g_{n+1}^{(k)} &= y_{n+1}^{(k)} + e^{x_{n+1}} \ , \quad k = 0, 1 \ . \end{aligned}$$

Assim, fazendo $n = 0$, obtemos:

$$P: \begin{cases} y_1^{(0)} = y_0 + h(f_0) = 1 + 0.1(0) = 1 \\ z_1^{(0)} = z_0 + h(g_0) = 0 + 0.1(1 + e^0) = 0.2 \end{cases}$$

$$E: \begin{cases} f_1^{(0)} = z_1^{(0)} = 0.2 \\ g_1^{(0)} = y_1^{(0)} + e^{x_1} = 1 + e^{(0.1)} = 2.1051 \end{cases}$$

$$C: \begin{cases} y_1^{(1)} = y_0 + \frac{h}{2}[f_0 + f_1^{(0)}] = 1 + \frac{0.1}{2}[0 + 0.2] = 1.01 \\ z_1^{(1)} = z_0 + \frac{h}{2}[g_0 + g_1^{(0)}] = 0 + \frac{0.1}{2}[2 + 2.1051] = 0.2053 \end{cases}$$

$$E: \begin{cases} f_1^{(1)} = z_1^{(1)} = 0.2053 \\ g_1^{(1)} = y_1^{(1)} + e^{x_1} = 1.01 + e^{(0.1)} = 2.1152 \end{cases}$$

Logo:

$$\begin{pmatrix} y_1^{(1)} \\ z_1^{(1)} \end{pmatrix} = \begin{pmatrix} 1.01 \\ 0.2053 \end{pmatrix} \simeq \begin{pmatrix} y(x_1) \\ z(x_1) \end{pmatrix} = \begin{pmatrix} y(0.1) \\ z(0.1) \end{pmatrix} .$$

Agora, fazendo $n = 1$, obtemos:

$$P: \begin{cases} y_2^{(0)} = y_1 + h(f_1) = 1.01 + 0.1(0.2053) = 1.0305 \\ z_2^{(0)} = z_1 + h(g_1) = 0.2053 + 0.1(2.1152) = 0.4168 \end{cases}$$

$$E: \begin{cases} f_2^{(0)} = z_2^{(0)} = 0.4168 \\ g_2^{(0)} = y_2^{(0)} + e^{x_2} = 1.0305 + e^{(0.2)} = 2.2519 \end{cases}$$

$$C: \begin{cases} y_2^{(1)} = y_1 + \frac{h}{2}[f_1 + f_2^{(0)}] = 1.01 + \frac{0.1}{2}[0.2053 + 0.4168] = 1.0411 \\ z_2^{(1)} = z_1 + \frac{h}{2}[g_1 + g_2^{(0)}] = 0.2053 + \frac{0.1}{2}[2.1152 + 2.2519] = 0.4237 \end{cases}$$

Logo:

$$\begin{pmatrix} y_2^{(1)} \\ z_2^{(1)} \end{pmatrix} = \begin{pmatrix} 1.0411 \\ 0.4237 \end{pmatrix} \simeq \begin{pmatrix} y(x_2) \\ z(x_2) \end{pmatrix} = \begin{pmatrix} y(0.2) \\ z(0.2) \end{pmatrix} .$$

Portanto, a solução do sistema dado é:

x_n	y_n	z_n
0.0	1	0
0.1	1.01	0.2053
0.2	1.0408	0.4231

Exemplo 12.17 - *Resolver o sistema dado no exemplo 12.15, usando o método de Euler Melhorado, fórmula (12.35).*

Solução: Para resolver um sistema usando (12.35), fazemos:

$$\begin{aligned} y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \ , \text{ e} \\ z_{n+1} &= z_n + \frac{h}{2}(\ell_1 + \ell_2) \ , \text{ onde:} \\ k_1 &= f(x_n, y_n, z_n) \ , \\ \ell_1 &= g(x_n, y_n, z_n) \ , \\ k_2 &= f(x_n + h, y_n + hk_1, z_n + hk_1) \ , \\ \ell_2 &= g(x_n + h, y_n + h\ell_1, z_n + h\ell_1) \end{aligned} \quad . \tag{12.44}$$

Fazendo então $n = 0$ em (12.44), e lembrando que: $f(x, y, z) = z$, $g(x, y, z) = y + e^x$, $x_0 = 0$, $y_0 =$

1, $z_0 = 0$, e $h = 0.1$, obtemos:

$$\begin{aligned} y_1 &= y_0 + \frac{h}{2}(k_1 + k_2) \ , \\ z_1 &= z_0 + \frac{h}{2}(\ell_1 + \ell_2) \ , \end{aligned}$$

onde:

$$\begin{aligned} k_1 &= f(x_0, y_0, z_0) = f(0, 1, 0) = 0 \ , \\ \ell_1 &= g(x_0, y_0, z_0) = g(0, 1, 0) = 2 \ , \\ k_2 &= f(x_0 + 0.1, y_0 + (0.1)k_1, z_0 + (0.1)k_1) = f(0.1, 1, 0) = 0 \ , \\ \ell_2 &= g(x_0 + 0.1, y_0 + (0.1)\ell_1, z_0 + (0.1)\ell_1) = g(0.1, 1.2, 0.2) = 2.3052 \ . \end{aligned}$$

Assim:

$$\begin{aligned} y_1 &= 1 + \frac{0.1}{2}(0 + 0) = 1 \ , \\ z_1 &= 0 + \frac{0.1}{2}(2 + 2.3052) = 0.2153 \ , \end{aligned}$$

Logo:

$$\begin{pmatrix} y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.2153 \end{pmatrix} \simeq \begin{pmatrix} y(x_1) \\ z(x_1) \end{pmatrix} = \begin{pmatrix} y(0.1) \\ z(0.1) \end{pmatrix} .$$

Fazendo agora $n = 1$ em 12.44), obtemos:

$$y_2 = y_1 + \frac{h}{2}(k_1 + k_2) \ ,$$

e

$$z_2 = z_1 + \frac{h}{2}(\ell_1 + \ell_2) \ ,$$

onde:

$$\begin{aligned} k_1 &= f(x_1, y_1, z_1) = f(0.1, 1, 0.2153) = 0.2153 \ , \\ \ell_1 &= g(x_1, y_1, z_1) = g(0.1, 1, 0.2153) = 2.1052 \ , \\ k_2 &= f(x_1 + 0.1, y_1 + (0.1)k_1, z_1 + (0.1)k_1) = f(0.2, 1.0215, 0.2368) = 0.2368 \ , \\ \ell_2 &= g(x_1 + 0.1, y_1 + (0.1)\ell_1, z_1 + (0.1)\ell_1) = g(0.2, 1.2105, 0.4258) \\ &= 1.2105 + e^{0.2} = 2.4319 \ . \end{aligned}$$

Assim:

$$\begin{aligned} y_2 &= 1 + \frac{0.1}{2}(0.2153 + 0.2368) = 1.0226 \ , \\ &\text{e} \\ z_1 &= 0.2153 + \frac{0.1}{2}(2.1052 + 2.4319) = 0.4422 \ , \end{aligned}$$

Logo:

$$\begin{pmatrix} y_2 \\ z_2 \end{pmatrix} = \begin{pmatrix} 1.0226 \\ 0.4422 \end{pmatrix} \simeq \begin{pmatrix} y(x_2) \\ z(x_2) \end{pmatrix} = \begin{pmatrix} y(0.2) \\ z(0.2) \end{pmatrix} .$$

Portanto, a solução do sistema dado é:

x_n	y_n	z_n
0.0	1	0
0.1	1	0.2153
0.2	1.0226	0.4422

Pelos exemplos desta seção, vemos que a aplicação de um método numérico a um sistema de equações ordinárias de primeira ordem se processa como no caso de uma única equação, só que aqui devemos aplicar o método numérico a cada uma das componentes do vetor.

12.6.2 Equações Diferenciais de Ordem Elevada

Finalmente, mostraremos como equações de ordem elevada podem ser escritas e portanto resolvidas como um sistema de equações de primeira ordem. Consideremos a equação diferencial de ordem n:

$$y^{(n)} = f(x, y, y', \ldots, y^{(n-1)}) \ .$$

com as condições iniciais:

$$y(x_0) = y_0 \ , \ y'(x_0) = y'_0 \ , \ \ldots \ , \ y^{(n-1)}(x_0) = y_0^{(n-1)} \ .$$

Novamente, para simplicidade, mas sem perda de generalidade, consideremos a equação diferencial de segunda ordem:

$$\begin{cases} y'' &= f(x, y, y') \\ y(x_0) &= y_0 \\ y'(x_0) &= y'_0 \end{cases} \qquad (12.45)$$

Podemos resolver qualquer equação diferencial de ordem elevada reduzindo-a a um sistema de equações diferenciais de primeira ordem. Para tanto basta fazer uma mudança adequada de variável, para a equação de segunda ordem, (12.45). Fazendo a seguinte mudança de variável:

$$y' = z \ ,$$

obtemos:

$$z' = g(x, y, z) \ .$$

Assim o **(p.v.i.)**, (12.45), se reduz a:

$$\begin{cases} y' = z \\ z' = g(x,y,z) \\ y(x_0) = y_0 \\ z(x_0) = z_0 \end{cases} \tag{12.46}$$

Observe que podemos reescrever o sistema (12.46) na forma:

$$\begin{cases} y' = f(x,y,z) \\ z' = g(x,y,z) \\ y(x_0) = y_0 \\ z(x_0) = z_0 \end{cases}$$

Daremos agora alguns exemplos.

Exemplo 12.18 - *Resolver a equação diferencial de segunda ordem:*

$$\begin{cases} y'' - y = e^x \\ y(0) = 1 \\ y'(0) = 0 \qquad x \in [0, 0.3] \ , \quad h = 0.1 \end{cases}$$

usando o método de Euler.

Solução: Fazendo $y' = z$ obtemos que: $z' = y + e^x$. Assim a equação de segunda ordem fica reduzida ao sistema:

$$\begin{cases} y' = z \\ z' = y + e^x \\ y(0) = 1 \\ z(0) = 0 \qquad x \in [0, 0.3] \ , \quad h = 0.1 \end{cases}$$

Observe que o sistema obtido é exatamente aquele dado no exemplo 12.15, e portanto já determinamos sua solução. Além disso, cabe salientar que, a solução aproximada da equação diferencial de segunda ordem encontra-se na primeira componente do vetor, isto é, apenas nos interessa o valor de y_n, apesar de termos de calcular, a cada passo, todas as componentes do vetor. Para a equação dada a solução exata é: $y(x) = \frac{1}{4}[e^x(1+2x) + 3e^{-x}]$.

Exemplo 12.19 - *Escrever a equação diferencial de terceira ordem:*

$$\begin{cases} y''' - 2xy'' + 4y' - x^2 y \ = \ 1 \\ y(0) = 1 \\ y'(0) = 2 \\ y''(0) = 3 \end{cases}$$

na forma de um sistema de equações diferenciais de primeira ordem.

Solução: Fazendo: $y' = z$ e $z' = w$, obtemos:

$$\begin{cases} y' \ = \ z \\ z' \ = \ w \\ w' \ = \ 2xw - 4z + x^2 + 1 \\ y(0) \ = \ 1 \\ z(0) \ = \ 2 \\ w(0) \ = \ 3 \end{cases}$$

Exercícios

12.18 - *Considere o sistema de equações diferenciais de primeira ordem:*

$$\begin{cases} y' &= y^2 - 2yz \\ z' &= xy + y^2 sen\ z \\ y(0) &= 1 \\ z(0) &= -1 \end{cases}$$

Resolva-o usando:

a) *método de Euler,*

b) *Método de Adams-Bashforth,*

c) *Método de Nystrom.*

12.19 - *Usando o par PC, onde:*

$$P: \quad y_{n+2} = y_{n+1} + \frac{h}{2}[-f_n + 3f_{n+1}].$$

$$C: \quad y_{n+2} = y_{n+1} + \frac{h}{12}[-f_n + 8f_{n+1} + 5f_{n+2}]$$

no modo $P(EC)E$, resolva a equação diferencial de segunda ordem:

$$\begin{cases} y'' + 3y' + 2y &= e^x \\ y(0) &= 1 \\ y(0) &= 2\ , \quad x \in [0, 0.4]\ , \quad h = 0.05. \end{cases}$$

12.7 Exercícios Complementares

12.20 - *Considere o seguinte problema de valor inicial:*

$$\begin{cases} y' = 2x^3 - 2xy \\ \\ y(0) = 1 \qquad x \in [0, 0.3]; \quad h = 0.15 \end{cases}$$

Rseolva-o:

a) *pelo método de Euler,*

b) *pelo método de Taylor de ordem 2.*

12.21 - *Resolva aproximadamente o problema de valor inicial:*

$$\begin{cases} y' = y + x^{\frac{3}{2}} \\ \\ y(0) = 1 \qquad x \in [0, 0.2]; \quad h = 0.1\ , \end{cases}$$

escolhendo q adequadamente tal que seja possível a aplicação do método de Taylor de ordem q.

12.22 - *O* **(p.v.i.)***:*

$$\begin{cases} y' = 3 \\ y(1) = 6 \quad \text{com} \quad h = 2 \, , \end{cases}$$

tem como solução exata $y(x) = 3x + 3$. *Usando o método de Euler, determine* $y(7)$. *Era de se esperar tal concordância mesmo com* h *bastante grande? Por que?*

12.23 - *Considere o método de Quade:*

$$y_{n+4} - \frac{8}{19}(y_{n+3} - y_{n+1}) - y_n = \frac{6}{19}h[f_{n+4} + 4(f_{n+3} + f_{n+1}) + f_n] \, .$$

a) *Determine sua ordem e a constante do erro.*

b) *Verifique se este método pode ser aplicado para resolver um* **(p.v.i.)** *com garantia de convergência.*

12.24 - *Mostre que a ordem do método de passo múltiplo:*

$$y_{n+2} + (b-1)y_{n+1}) - by_n = \frac{1}{4}h[(b+3)f_{n+1} + (3b+1)f_n] \, .$$

é 2 se $b \neq -1$ *e 3 se* $b = -1$. *Mostre que o método não é estável se* $b = -1$.

12.25 - *Considere o método explícito de 2-passos:*

$$y_{n+2} + 4y_{n+1} - 5y_n = h[b_1 f_{n+1} + b_0 f_n] \, .$$

a) *Mostre que* b_1 *e* b_0 *podem ser determinados se a ordem do método for* $q = 3$.

b) *Calcule a constante do erro.*

c) *Este método de ordem 3 foi aplicado ao* **(p.v.i.)***:*

$$\begin{cases} y' = y \\ y(0) = 1 \qquad x \in [0, 1.0]; \quad h = 0.1 \, , \end{cases}$$

e os cálculos foram realizados com seis casas decimais. O valor de y_{10} *tornou-se negativo. Explique porque o erro foi tão grande.*

12.26 - *Resolva o seguinte problema de valor inicial:*

$$\begin{cases} y' = xy - y^2 + 1 \\ y(0) = 1 \qquad x \in [0, 0.2]; \quad h = 0.05 \end{cases}$$

usando:

a) *método de Euler,*

b) *método previsor-corretor, onde*

$$\begin{cases} P : y_{n+1} = y_n + hf_n \\ C : y_{n+1} = y_n + \frac{h}{2}[f_n + f_{n+1}] \end{cases}$$

usando o par PC *no modo* $P(EC)^2E$,

c) *método previsor-corretor, onde*

$$\begin{cases} P: y_{n+2} = y_{n+1} + \frac{h}{2}[-f_n + 3f_{n+1}] \\ C: y_{n+2} = y_{n+1} + \frac{h}{12}[-f_n + 8f_{n+1} + 5f_{n+2}] \end{cases}$$

usando o par PC no modo $P(EC)E$. Obtenha y_1 pelo método de Taylor de ordem 3 ou pelo método de Heun.

12.27 - *Prove que se o método de Runge-Kutta é consistente então $\sum_{r=1}^{R} c_r = 1$.*

12.28 - *Resolva o seguinte sistema de equações diferenciais ordinárias:*

$$\begin{cases} y' = y + z \\ z' = 2y + 3z \\ y(0) = 2 \\ z(0) = 0 \qquad x \in [0, 0.3]; \quad h = 0.1 \end{cases}$$

usando os métodos do exercício 12.26.

12.29 - *Dado o* **(p.v.i.)***:*

$$\begin{cases} 2yy'' - 4xy^2 + 2(sen x)y^4 = 6 \\ y(1) = 0 \\ y'(1) = 15 \end{cases}$$

a) *reduza-o a um sistema de equações de primeira ordem,*

b) *resolva o* **(p.v.i.)** *dado, usando um método de Runge-Kutta de ordem 2, a sua escolha.*

12.30 - *Resolva o problema de valor inicial de terceira ordem:*

$$\begin{cases} xy''' - x^2y'' + (y')^2y = 0 \\ y(0) = 1 \\ y'(0) = 2 \\ y'''(0) = 3 \end{cases}$$

usando o método previsor dado por (12.24), no modo $P(EC)E$.

12.31 - *Resolva o problema de valor inicial de segunda ordem:*

$$\begin{cases} y'' - 3y' + 2y = 0 \\ y(0) = -1 \\ y'(0) = 0 \qquad x \in [0, 0.3]; \quad h = 0.1 \end{cases}$$

usando os métodos do exercício 12.26.

12.8 Problemas Aplicados e Projetos

12.1 - *Um projétil é lançado para o alto contra a resistência do ar. Suponha que a equação do movimento é dada por:*

$$\frac{dv}{dt} = -32 - \frac{cv^2}{m} \ .$$

Se $\frac{c}{m} = 2$ *e* $v(0) = 1$, *determine o tempo necessário para que o projétil alcance sua altura máxima, resolvendo o problema de valor inicial pelo método previsor-corretor:*

$$\begin{cases} P : y_{n+2} = y_{n+1} + \frac{h}{2}[-f_n + 3f_{n+1}] \\ \\ C : y_{n+2} = y_n + \frac{h}{3}[f_n + 4f_{n+1} + f_{n+2}] \end{cases}$$

no modo $P(EC)E$. *Obtenha os valores iniciais necessários pelo método de Taylor se ordem 2, ou pelo método de Heun. Considere* $h = 0.01$.

12.2 - *Um corpo com uma massa inicial de* $200kg$ *é acelerado por uma força constante de* $200N$. *A massa decresce a uma taxa de 1 kg/s. Se o corpo está em repouso em* $t = 0$ *encontre sua velocidade ao final de* $50s$. *Sabe-se que a equação diferencial é dada por:*

$$\frac{dv}{dt} \ = \ \frac{2000}{200 - t} \ .$$

Resolva o problema usando:

a) *método de Euler,*

b) *método de Euler melhorado,*

c) *método de Nystrom.*

Observação: A solução exata desse **(p.v.i.)** é:

$$v \ = \ 2000 log \left[\frac{200}{200 - t} \right] \ ,$$

de modo que $v(50) = 575.36$.

12.3 - *Suponha que o corpo descrito no problema 12.2 está sujeito a uma resistência do ar igual a duas vezes a velocidade. A equação diferencial agora é:*

$$\frac{dv}{dt} \ = \ \frac{2000 - 2v}{200 - t} \ .$$

Se o corpo está em repouso em $t = 0$, *a solução exata é* $v = 10t - 40t^2$, *de modo que* $v(50) = 437.50$. *Resolva o problema numericamente usando um método de Runge-Kutta de ordem 4.*

12.4 - *Considere o conjunto massa-mola dado na Figura 12.2.*

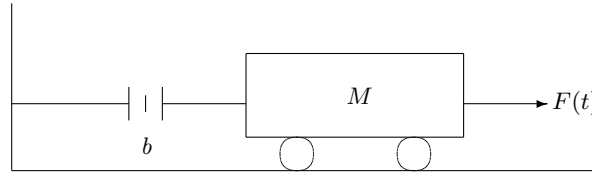



Figura 12.2

A equação diferencial que descreve o sistema é:

$$M\frac{dv(t)}{dt} + bv(t) = F(t) \ ,$$

onde $v(t)$ é a velocidade no instante $t > 0$. Assuma que:

$$t_i = ih, \quad i = 0, 1, \ldots 5; \qquad h = 0.4 \ ;$$

$$v(0) = 0m/s, \quad b = 3Kg/s, \quad M = 1Kg, \quad F(t) = 1N \ .$$

a) *Calcule $v(2)$ pelo método de Euler.*

b) *Calcule $v(2)$ pelo método de Taylor de ordem 2.*

c) *Calcule $v(2)$ pelo método previsor-corretor dado no item* **b)** *do exercício 12.26.*

d) *Sabendo que a solução exata é dada por:*

$$v(t) = \frac{e^{-3t}}{9} + \frac{t}{3} - \frac{1}{9} \ ,$$

faça uma tabela comparando os resultados obtidos nos itens **a), b)** *e* **c)** *com a solução exata.*

12.5 - *A taxa de fluxo de calor entre dois pontos de um cilindro aquecido em uma das extremidades é dada por:*

$$\frac{dQ}{dt} \ = \ \lambda \ A\frac{dT}{dt} \ ,$$

onde λ é uma constante, A é a área da seção reta do cilindro, Q é o fluxo de calor, T é a temperatura, t é o tempo, e x é a distância até a extremidade aquecida. Como a equação envolve duas derivadas, podemos simplificar a equação dada assumindo que:

$$\frac{dT}{dt} \ = \ \frac{100(L-x)(20-t)}{100-xt} \ ,$$

onde L é o comprimento do cilindro. Combinando as duas equações, obtemos:

$$\frac{dQ}{dt} \ = \ \lambda \ A\frac{100(L-x)(20-t)}{100-xt} \ .$$

Utilizando o método de Runge-Kutta de quarta ordem, calcule o fluxo de calor de $t = 0s$ até $t = 2s$ em passos de $0.01s$. Adote $\lambda = 0.3cal.cm/sC^{o}$, $A = 10cm^2$, $L = 20cm$, $x = 2.5cm$, e como condição inicial, $Q = 0$ em $t = 0$.

12.6 - *Um paraquedista em queda livre está sujeito à seguinte equação:*

$$\frac{dv}{dt} = g - c\,\frac{c}{m}v \ ,$$

onde m *é a massa do paraquedista,* g *é a aceleração da gravidade,*

Supondo que no instante $t_0 = 0$*, o paraquedista salte de um avião com velocidade vertical* $v_0 = 0$*, determine, usando o método de Euler com passo de* $0.01s$*, a velocidade do paraquedista nos instantes de tempo:* $2, 4, 6$ *e* $8s$ *em* m/s*.*

12.7 - *Numa reação química, uma molécula de um reagente* A *combina-se com uma molécula de um outro reagente* B *para formar uma molécula de um produto* C*. Sabe-se que a concentração* $y(t)$ *de* C*, no tempo, é solução do seguinte* **(p.v.i.)***:*

$$\begin{cases} y' = k(a-y)(b-y) \\ \\ y(0) = 0 \end{cases}$$

onde k *é a constante de reação,* a *e* b *são, respectivamente, a concentração inicial do reagente* A *e* B*. Considerando os seguintes dados:* $k = 0.01, a = 70 milimoles/litro$ *e* $b = 50 milimoles/litro$*, determine a concentração do produto* C *sobre o intervalo* $[0, 20]$*, usando o método de Heun, com* $h = 0.5$*, e compare os resultados obtidos com a solução exata:*

$$y(t) = \frac{350(1 - e^{-0.2t})}{7 - 5e^{-0.2t}} \ .$$

12.8 - *Para um circuito simples RL, a lei de voltagem de Kirchoff, exige (se a lei de Ohm vale), que:*

$$L\frac{di}{dt} + Ri = 0 \ ,$$

onde i *é a corrente,* L *é a indutância, e* R *é a resistência. Sabendo que* $i(0) = 10^{-3}$ *e considerando que* $L = R = 1$*, resolva o problema usando um método numérico à sua escolha, obtendo a solução com precisão de* 10^{-4}*.*

12.9 - *Em contraste com o problema anterior, resistores reais podem não obedecer a lei de Ohm. Por exemplo, a queda de voltagem pode ser não linear e a dinâmica do circuito é descrita por:*

$$L\frac{di}{dt} + \left(\frac{-i}{I} + \left(\frac{i}{I}\right)^3\right) R = 0 \ ,$$

onde todos os parâmetros são definidos como no problema anterior, e I *é a conhecida corrente de referência. Considere* $I = 1$*. Resolva então o problema para* i *como função do tempo usando as mesmas condições especificadas no problema anterior.*

12.10 - *O estudo sobre o crescimento da população é importante para uma variedade de problemas de engenharia. Um modelo simples, onde o crescimento está sujeito à hipótese de que a taxa de variação da população* p *é proporcional à população num instante* t*, isto é:*

$$\frac{dp}{dt} = Gp \ ,$$

onde G *é a taxa de variação (por ano). Suponha que no instante* $t = 0$*, uma ilha tem uma população de* 10000 *habitantes. Se* $G = 0.075$ *por ano, utilize o método de Euler para prever a população em* $t = 20$ *anos, usando comprimento de passo de* 0.5 *ano.*

12.11 - *A resposta $y(>0)$ de uma certa válvula hidráulica sujeita a uma entrada de variação senoidal é dada por:*

$$\frac{dy}{dt} = \sqrt{2\left(1 - \frac{y^2}{sen^2 t}\right)}, \quad \text{com} \quad y_0 = 0, t_0 = 0 \ .$$

Deseja-se obter a solução numérica desse **(p.v.i.)** *usando um método numérico de ordem 3. Pergunta-se: o algoritmo de Taylor pode ser usado para obter a solução desse* **(p.v.i.)**? *Se possível resolva-o pelo algoritmo de Taylor de ordem 3, caso contrário use o método de Runge-Kutta de ordem 3.*

12.12 - *A corrente i num circuito LR num instante t qualquer depois que uma chave é ligada em $t = 0$ pode ser expressa pela equação:*

$$\frac{di}{dt} = \frac{(Esen\ \omega t - R)}{L} \ ,$$

onde $E = 50$ Volts, $L = 1$ Henry, $\omega = 300$, $R = 50$ Ohms e a condição inicial é que $i(0) = 0$. Resolva numericamente o **(p.v.i.)** *por um método de Runge-Kutta de ordem 3 e compare sua solução com a solução analítica:*

$$i = \frac{E}{Z^2}(Rsen\ \omega t - \omega Lcos\ \omega t + \omega L e^{-Rt/L} \ ,$$

onde $Z = \sqrt{R^2 + \omega^2 L^2}$.

12.13 - *Uma quantidade de 10 kg de um certo material é lançado em um recipiente contendo 60 kg de água. A concentração da solução, c, em percentagem, a qualquer instante t é expressa como:*

$$(60 - 1.212c)\frac{dc}{dt} = \frac{k}{3}(200 - 14c)(100 - 4c) \ ,$$

onde k, o coeficiente de transferência de massa, é igual a 0.0589. A condição incial é que em $t = 0, c = 0$. Determine a relação entre c e t usando um método de Runge-Kutta de ordem 2.

12.14 - *A equação de Van der Pol, da eletrônica, é:*

$$y'' + (1 - y^2)y' + y = 0 \ ,$$

com condições iniciais: $y(0) = 0.5$ e $y'(0) = 0$. Obtenha o valor de y, y', y'' em $t = 0.4$ usando um método de Runge-Kutta de ordem 2.

12.15 - *Um sistema simples em vibração tem uma massa sujeita ao atrito de Coulomb, de modo que a equação de seu movimento é:*

$$my'' + n^2 y = \begin{cases} -A, & y' > 0 \\ +A, & y' < 0 \end{cases}$$

onde A é uma constante, e $y(0) = 3$, $y'(0) = 0$. Considerando $m = 1, n = 0.8$ e $A = 2$ e usando um método de Runge-Kutta de ordem 3, obtenha o valor de y em $t = 5$ segundos, com precisão de 10^{-2}.

12.16 - *Engenheiros ambientais e biomédicos precisam frequentemente prever o resultado de uma relação predador-presa ou hospedeiro-parasita. Um modelo simples para esse tipo de relação é dado pelas equações de Lotka-Volterra:*

$$\frac{dH}{dt} = g_1 H - d_1 PH \ ,$$

$$\frac{dP}{dt} = -d_2 P + g_2 PH \ .$$

onde H e P são, respectivamente, por exemplo, o número de hospedeiros e parasitas presentes. As constantes d e g representam as taxas de mortalidade e crescimento, respectivamente. O índice 1 refere-se ao hospedeiro e o índice 2 ao parasita. Observe que essas equações formam um sistema de equações acopladas. Sabendo que no instante $t_0 = 0$, os valores de P e H são , respectivamente, 5 e 20, e que $g_1 = 1, d_1 = 0.1, g_2 = 0.02$ e $d_2 = 0.5$, utilize um método numérico para calcular os valores de H e P de 0 até $2s$, usando passo de $0.01s$.

12.17 - *As equações:*

$$\begin{cases} y'(t) = -\dfrac{2y}{\sqrt{y^2+z^2}} \\ z'(t) = 1 - \dfrac{2z}{\sqrt{y^2+z^2}} \end{cases}$$

descrevem a trajetória de um pato nadando em um rio e tentando firmemente chegar à posição t. Veja Figura 12.4.

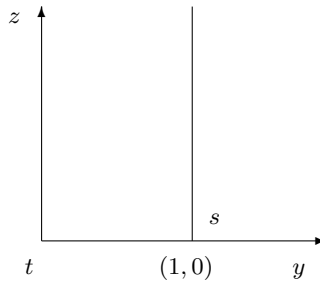



Figura 12.4

O pato parte de s, de modo que $y(0) = 1$ e $z(0) = 0$.

Calcule a trajetória do pato até $t = 0.2$ usando o par PC no modo $P(EC)^2$, onde o previsor é o método de Euler e o corretor a regra do trapézio. Considere $h = 0.1, 0.01, 0.001$. Quantas casas decimais você pode garantir que estão corretas?

Capítulo 13

Solução Numérica de Equações Diferenciais Parciais

13.1 Introdução

Um modelo matemático é uma representação, idealizada e muitas vezes simplificada, da natureza. Quando derivado de maneira criteriosa, sua solução simula propriedades dos processos naturais envolvidos, tais como velocidade e pressão no escoamento de um fluido, temperatura na previsão do tempo, trajetória de um satélite artificial, etc... Assim, as soluções das equações de um modelo, devem apresentar um comportamento compatível com as propriedades do problema modelado, não sendo possível, na maioria dos casos, justificar a utilização de hipóteses simplificadoras (tal como linearidade) que permitam a determinação de uma solução exata. Daí a necessidade da procura de soluções numéricas, ou aproximadas.

A importância da modelagem matemática cresceu muito nos ultimos anos porque a análise detalhada de um modelo e suas propriedades permite um melhor entendimento do evento modelado e, mais do que isso, permite a simulação de mudanças nos parâmetros do modelo e a respectiva análise da respectiva resposta, que não seriam possíveis na situação real. Por exemplo, no projeto de um carro, alterações na forma resultam em modificações no comportamento aerodinâmico, de cuja análise obtem-se um indicador dos ganhos e/ou perdas em performance; a localização ideal da asa de um avião em relação ao casco pode ser obtida pela resposta à simulação matemática das equações da aerodinâmica, e até a melhor política de vacinação contra doenças transmissíveis, tipo sarampo, podem ser decididos com base em modelos matemáticos. A economia de tempo gerada por esta maneira de projetar um produto, ou tomar decisões, é clara, diminuindo sensivelmente o número de protótipos ou modelos em tamanho reduzido a serem construídos e ensaiados. No projeto de equipamentos complexos, o mínimo que a simulação através da modelagem matemática coopera é na eliminação de casos triviais ou impossíveis, fornecendo um guia seguro para a seleção dos casos a serem ensaiados em modelos de escala reduzida ou para a construção de protótipos.

Como virtualmente todas as áreas em matemática aplicada o principal objetivo das equaçẽs diferenciais é o de modelar matematicamente eventos da natureza nas mais diversas áreas de aplicação. Um modelo com várias variáveis independentes envolve necessariamente derivadas com respeito a cada umas dessas variáveis ou derivadas parciais, sendo portanto denominado uma *equação diferencial parcial* ou um *sistema de equações diferenciais parciais*.

Resumindo as idéias apresentadas acima o processo de simulação é constituído de três partes nitida-

mente distintas: A fase de modelagem, isto é a construção de um conjunto de equações matemáticas que reputamos representar os fenômenos e os processos modelados. A segunda fase de solução desse conjunto de equações, normalmente utilizando técnicas de discretização numérica e um computador e finalmente a fase de interpretação dos resultados face às características do problema original. Esse é um processo complexo que exige do profissional um conjunto bastante amplo de habilidades; exige um bom conhecimento de engenharia para os ajustes finos do modelo desprezando complicações que não são fundamentais, um bom conhecimento de métodos numéricos para selecionar aquele que melhor adapta-se ao problema e finalmente um bom faro de detetive para analisar os resultados e interpretá-los à luz das restrições e características do problema. Este capítulo trata exclusivamente da segunda fase desse processo.

É claro portanto que a solução do modelo matemático, ou seja, das equações representantes desse modelo, é fundamental para a compreensão do fenômeno modelado. Este é o papel da discretização das equações parciais, uma vez que, como já dissemos, uma solução analítica nem sempre está disponível; ou ela tem uma forma não prática ou é impossível de ser obtida. Assim os métodos numéricos são amplamente usados para a aproximação dessas soluções. A essência dos métodos numéricos está na representação discreta (finita) do problema que, em geral, é originalmente modelado como um contínuo. Essa discretização é que viabiliza o uso dos computadores no tratamento numérico das equações diferenciais.

O objetivo deste capítulo é apresentar uma introducao à solução numérica de equações diferenciais parciais, através da discretização por diferenças finitas, enfatizando os principais conceitos com vistas às aplicações práticas. Nos restringiremos aos casos das equações parabólicas e elípticas, por serem mais adequados para a solução por diferenças finitas. No entanto, equações hiperbólicas também podem ser resolvidas por diferenças finitas, mas nesse caso temos que ser mais cuidadosos devido ao aparecimento de singularidades nas soluções.

Por razões práticas e talvez também didáticas é costume na literatura classificar as equações diferenciais parciais em tres grupos distintos: Equações Parabólicas, Equações Elípticas e Equações Hiperbólicas. No caso de equações de segunda ordem em duas dimensões da forma:

$$a(x,y)u_{xx} + b(x,y)u_{xy} + c(x,y)y_{yy} + d(x,y,u_x,u_y,u) \;=\; 0 \;, \tag{13.1}$$

onde u_x denota a derivada de u em relação à variàvel x e a, b, c e d são funções conhecidas, a classificação é feita em função do sinal do discriminante: $\Delta = b^2 - 4ac$.

1. $\Delta = 0$ - Equação Parabólica,
2. $\Delta < 0$ - Equação Elíptica,
3. $\Delta > 0$ - Equação Hiperbólica.

É claro que, como, a, b e c dependem de x, y, Δ também depende e portanto o sinal de Δ pode variar para diferentes valores de x e y, e nesse caso a equação muda de tipo no domínio de definição. De forma que é perfeitamente possível que uma mesma equação seja de um, dois ou mais tipos dependendo da região do domínio considerada.

A classificação acima pode em princípio parecer irrelevante, mas de fato ela é de extrema importância, tanto do ponto de vista prático das aplicações quanto do ponto de vista da solução numérica. Na área de aplicações temos, por exemplo, que as equações elípticas são adequadas para modelar problemas de equilíbrio, as equações parabólicas problemas de difusão e as equações hiperbólicas problemas de convecção. Portanto a classificação constitui-se em um teste da adequação do modelo ao problema. Já

no contexto de solução numérica, sabemos da teoria matemática das equações diferenciais parciais, que as equações parabólicas e elípticas apresentam soluções altamente regulares, enquanto as equações hiperbólicas podem apresentar soluções singulares. Essa informação pode ser crucial no desenvolvimento de um método numérico.

13.2 Equações Parabólicas

Nesta seção apresentamos os métodos mais conhecidos para solução de equações parabólicas. Essas equações aparecem no modelamento de processos conhecidos como de difusão. Por exemplo, a distribuição de temperatura em uma barra de metal cujas extremidades são mantidas a temperaturas conhecidas e termicamente isolada ao longo do comprimento, veja figura, é descrita pela equação do calor:

$$u_t = \alpha u_{xx} . \tag{13.2}$$

COLOCAR FFIGURA 13.1

Supomos que a barra tem comprimento L, coincide com o eixo x e tem uma das extremidades localizada na origem do eixo x. Portanto a temperatura $u(x,t)$, além da equação (13.2), deve satisfazer as seguintes condições de fronteira:

$$u(0,t) = T_a , \quad u(L,t) = T_b , \quad t \geq 0 . \tag{13.3}$$

A equação (13.2) justifica-se assumindo que a barra é feita de um material homogêneo, de forma que a temperatura numa secção transversal ao eixo x seja constante. Podemos então considerá-la como unidimensional e portanto a distribuição de temperatura será uma função da posição x e do tempo t. É claro que se quisermos levar a espessura da barra em consideração basta introduzir mais uma variável e teremos então a equação do calor em duas dimensões:

$$u_t = \alpha(u_{xx} + u_{yy}) . \tag{13.4}$$

que não apresenta muito mais dificuldades de solução do que a versão unidimensional. Obviamente as condições de fronteira devem ser modificadas para adequarem-se à geometria do problema.

Voltando ao problema unidimensional note que a distribuição de calor ao longo da barra depende claramente de qual é a temperatura no instante inicial em que começamos nossa observação. Isto é traduzido matematicamente em uma condição, chamada *condição inicial* que normalmente expressamos da forma:

$$u(x,0) = \psi(x) , \quad 0 \leq x \leq L , \tag{13.5}$$

onde $\psi(x)$ é uma função conhecida.

A equação (13.2), juntamente com a condição inicial (13.5) e as condições de fronteira (13.3) formam um problema de equações diferenciais parciais para o qual buscamos uma solução. O leitor poderá

facilmente constatar consultanto um livro de matemática aplicada, por exemplo (citar o livro do Djairo), ou ([**?**], página 100) que a solução do problema acima é:

$$u(x,t) = T_a + (T_b - T_a)\frac{x}{L} + \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{T_b\cos(n\pi) - T_a}{n} exp\left(-\frac{\alpha n^2\pi^2 t}{L^2}\right)\sin\left(\frac{n\pi x}{L}\right) \quad (13.6)$$

$$+ \sum_{n=1}^{\infty} C_n \exp\left(-\frac{\alpha n^2\pi^2 t}{L^2}\right)\sin\left(\frac{n\pi x}{L}\right) , \quad (13.7)$$

onde

$$C_n = \frac{2}{L}\int_0^L \psi(x)\sin\left(\frac{n\pi x}{L}\right) dx .$$

O leitor deve estar se perguntando, se temos a solução analítica do nosso problema o que mais resta a fazer? A resposta a esta pergunta é: de fato nosso problema modelo permite uma solução analítica em forma de série de potências, por ser um problema muito simples. Mas este não será sempre o caso. Existem problemas para os quais uma solução analítica não é possível. Para dar um exemplo basta fazer α em (13.2) uma função de x ou t e as coisas se complicam consideravelmente. Só uma solução numérica será possível nesse caso.

Na discução acima procuramos, de maneira um tanto concisa, situar o leitor diante do problema de modelagem e a necessidade de solução numérica da equação diferencial resultante. Passaremos a seguir a discutir as técnicas numéricas que permitem a obtenção de uma solução. Com esse objetivo introduzimos brevemente a aproximação das derivadas de uma função de uma variável, por diferenças finitas.

Diferenças Finitas

A idéia geral do método de diferenças finitas é a discretização do domínio e a substituição das derivadas presentes na equação diferencial por aproximações envolvendo somente valores numéricos da função. Na prática substitui-se as derivadas pela razão incremental que converge para o valor da derivada quando o incremento tende a zero. Dizemos então que o problema foi *discretizado*. Quando o domínio tem mais de uma variável, a idéia acima é aplicada para cada uma delas separadamente.

Seja x_0 um número real pertencente ao domínio em questão e h um número positivo. Definimos *malha* de passo h associada a x_0 como o conjunto de pontos

$$x_i = x_0 \pm ih , \qquad i = 1, 2, \ldots, N .$$

Nos pontos dessa malha serão calculadas aproximações de uma função $y(x)$ e suas derivadas.

A ferramenta matemática básica no cálculo de aproximações para as derivadas é a série de Taylor que relaciona valores da função e suas derivadas num ponto x com valores dessa mesma função numa vizinhaça de x, ou seja com $y(x+h)$. Se $y(x)$ tem derivadas até a ordem $n+1$ em x, obtemos:

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!}y''(x) + \cdots + \frac{h^n}{n!}y^{(n)}(x) + \quad (13.8)$$

$$+ \frac{h^{n+1}}{(n+1)!}y^{(n+1)}(\xi) , \qquad x < \xi < x+h , \quad (13.9)$$

o último termo da expressão acima representa o erro da aproximação de $y(x+h)$ pelo polinômio (na variável h) de grau n.

Se $n = 1$ em (13.9) teremos a *fórmula progressiva* que utiliza a *diferença progressiva* ($\Delta y(x)$) e seu erro, ou seja:

$$\begin{aligned} y'(x) &= \frac{y(x+h)-y(x)}{h} - \frac{h}{2}y''(\xi) \\ &= \frac{1}{h}\Delta y(x) - \frac{h}{2}y''(\xi)\ . \end{aligned}$$

De modo semelhante, tomando $-h$ em (13.9), ainda com $n = 1$, obtemos a *fórmula regressiva* que utiliza a *diferença regressiva* ($\nabla y(x)$) e seu erro, ou seja:

$$\begin{aligned} y'(x) &= \frac{y(x)-y(x-h)}{h} + \frac{h}{2}y''(\xi) \\ &= \frac{1}{h}\nabla y(x) + \frac{h}{2}y''(\xi)\ . \end{aligned}$$

Fazendo agora, $n = 2$ em (13.9) com h e $-h$, respectivamente, temos

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!}y''(x) + \frac{h^3}{3!}y'''(\xi_2)$$

e

$$y(x-h) = y(x) - hy'(x) + \frac{h^2}{2!}y''(x) - \frac{h^3}{3!}y'''(\xi_1)\ .$$

Subtraindo a última expressão da penúltima obtemos a *fórmula centrada* que utiliza a *diferença central* ($\delta_h y(x)$) e seu erro, ou seja:

$$\begin{aligned} y'(x) &= \frac{y(x+h)-y(x-h)}{2h} + \frac{h^2}{3!}y'''(\xi) \\ &= \frac{1}{2h}\delta_h y(x) + \frac{h^2}{3!}y'''(\xi). \end{aligned}$$

onde $\xi \in (x-h, x+h)$.

Quando se fizer necessário deixar claro o passo h para o qual um operador está sendo utilizado escreveremos: $\delta_h, \delta_h^2, \Delta_h, \nabla_h$, caso contrário o indice h será omitido.

Erro e Ordem de Aproximação de Uma Fórmula de Diferença

Seja $\mathcal{F}(x)$ uma fórmula de diferença para aproximação da derivada de ordem q de uma função $y(x)$ com erro $\mathcal{E}(x)$. Então:

$$y^{(q)}(x) = \mathcal{F}(x) + \mathcal{E}(x).$$

Dizemos que a fórmula $\mathcal{F}(x)$ é de ordem p se $\mathcal{E}(x) = h^p\mathcal{R}(x)$, onde $\mathcal{R}(x)$ não depende de h. Nesse caso usamos a notação $\mathcal{E}(x) = O(h^p)$. Essa notação significa que $\lim_{h\to 0} \dfrac{\mathcal{R}(x)}{h^p}$ é uma constante finita.

Por exemplo, no caso da fórmula centrada temos que:

$$\mathcal{F}(x) = \frac{y(x+h) - y(x-h)}{2h} \quad \text{e} \quad \mathcal{E}(x) = \frac{h^2}{3!} y'''(\xi)$$

de forma que essa fórmula é de segunda ordem.

Seguindo as mesmas idéias podemos estabelecer uma expressão para o cálculo aproximado da segunda derivada.Tomando $n = 3$ em (13.9) com h e $-h$ obtemos:

$$y(x+h) = y(x) + hy'(x) + \frac{h^2}{2!} y''(x) + \frac{h^3}{3!} y'''(x) + \frac{h^4}{4!} y^{(4)}(\xi_1) ,$$

e

$$y(x-h) = y(x) - hy'(x) + \frac{h^2}{2!} y''(x) - \frac{h^3}{3!} y'''(x) + \frac{h^4}{4!} y^{(4)}(\xi_2) .$$

Somando estas duas últimas expressões e isolando $y''(x)$ temos:

$$y''(x) = \frac{y(x+h) - 2y(x) + y(x-h)}{h^2} + \frac{h^2}{12} 1h^2 \delta^2_{\frac{h}{2}} y(x) + \frac{h^2}{12} y^{(4)}(\xi) ,$$

onde $\xi \in (x-h, x+h)$.

As fórmulas de diferenças finitas em uma dimensão obtidas acima, podem agora ser utilizadas em cada variável para gerar aproximações para as derivadas parciais de uma função de várias variáveis. Assim, temos as seguintes fórmulas:

progressiva

$$\begin{aligned} u_t(x,t) &= \frac{u(x,t+k) - u(x,t)}{k} - \frac{k}{2} u_{tt}(x,\zeta) \\ &= \frac{1}{k} \Delta_t u(x,t) - \frac{k}{2} u_{tt}(x,\zeta) , \qquad (t < \zeta < t+k) . \end{aligned} \tag{13.10}$$

regressiva

$$\begin{aligned} u_t(x,t) &= \frac{u(x,t) - u(x,t-k)}{k} + \frac{k}{2} u_{tt}(x,\zeta) \\ &= \frac{1}{k} \nabla_t u(x,t) + \frac{k}{2} u_{tt}(x,\zeta) , \qquad (t-k < \zeta < t) . \end{aligned} \tag{13.11}$$

central

$$\begin{aligned} u_x(x,t) &= \frac{u(x+h,t)-u(x-h,t)}{2h} - \frac{h^2}{6}u_{xxx}(\xi,t) \\ &= \frac{1}{2h}\delta_x u(x,t) - \frac{h^2}{6}u_{xxx}(\xi,t)\ , \qquad (x-h<\xi<x+h)\ . \end{aligned}$$

$$\begin{aligned} u_{xx}(x,t) &= \frac{u(x+h,t)-2u(x,t)+u(x-h,t)}{h^2} - \frac{h^2}{12} \\ &= \frac{1}{h^2}\delta_x^2 u(x,t) - \frac{h^2}{12}u_{xxx}(\xi,t)\ , \qquad (x-h<\xi<x+h)\ . \end{aligned} \tag{13.12}$$

$$\begin{aligned} u_{tt}(x,t) &= \frac{u(x,t+k)-2u(x,t)+u(x,t-k)}{k^2} - \frac{k^2}{12} \\ &= \frac{1}{k^2}\delta_t^2 u(x,t) - \frac{k^2}{12}u_{ttt}(x,\zeta)\ , \qquad (t-h<\zeta<t+h)\ . \end{aligned}$$

$$\begin{aligned} u_{xt}(x,t) &= \frac{u(x+h,t+k)-u(x+h,t-k)-u(x-h,t+k)+u(x-h,t-k)}{4hk} \\ &\quad -\frac{h^2}{6}u_{xxxt}(\xi_1,\zeta_1) - \frac{k^2}{} \\ &= O(h^2+k^2)\ , \quad x-h<\xi_1,\xi_2<x+h, \quad \text{e} \quad t-k<\zeta_1,\zeta_2<t+k\ . \end{aligned}$$

onde denotamos por $u_t(x,t)$ a derivada parcial da função u com relação a t e por $\Delta_t u(x,t)$ a diferença progressiva de u na variável t.

O Problema de Dirichlet

Em aplicações, uma forma mais comum de ocorrência de equações parabólicas é a forma:

$$\begin{aligned} u_t - g(x,t)u_{xx} &= r(x,t,u,u_x)\ , \quad a<x<b \ \text{ e } \ 0<t<T\ , \\ u(x,0) &= \psi(x)\ , \qquad a\le x\le b\ , \\ u(a,t) &= f(t)\ , \qquad 0<t<T\ , \\ u(b,t) &= g(t)\ , \qquad 0<t<T\ . \end{aligned} \tag{13.13}$$

que é chamado de **Problema de Dirichlet.**

Teorema 13.1 - *Se $g(x,t)$ é contínua e limitada e $r(x,t,u,u_x)$ é monotonicamente decrescente em u, então existe uma única solução para (13.13).*

Prova: A prova deste teorema pode ser encontrada em ....

Para mais detalhes sobre a teoria geral de equações parabólicas ver Friedman [**?**] e Bernstein [**?**].

O modelo fundamental das equações parabólicas é a equação do calor:

$$u_t - \alpha(x,t)u_{xx} \ = \ r(x,t) \ , \qquad \alpha(x,t) > 0 \ , \quad 0 \leq x \leq L \ , \quad 0 < t < T \ , \tag{13.14}$$

$$u(x,0) = \psi(x), \quad 0 \leq x \leq L \quad \text{condição inicial} \ ,$$

$$\left.\begin{array}{l} u(0,t) = f(t) \ , \quad 0 < t < T \ , \\ u(L,t) = g(t) \ , \quad 0 < t < T \ , \end{array}\right\} \text{condições de fronteira.}$$

É importante lembrar que no caso das equações parabólicas, a solução num ponto interior depende de toda a condição inicial. Além disso, o operador parabólico regulariza a condição inicial, ou seja, num ponto interior ($t > 0$), mesmo que bem próximo da condição inicial ($t = 0$), a solução é completamente suave (regular), mesmo que, por exemplo $f(0) \neq \psi(0)$. A mesma observação é válida próximo de $x = L$.

Discretização

As técnicas de solução de equações parabólicas serão apresentadas para o problema modelo da equação do calor, dado por (13.14), com $\alpha(x,t)$ constante e $r(x,t) \ \equiv \ 0$.

Consideremos, então, o problema de Dirichlet dado por (13.14) e as condições iniciais e de fronteira. Dividindo o intervalo $[0, L]$, da variável espacial x, em N partes iguais de comprimento h, temos os $N+1$ pontos $x_i = ih, i = 0, 1, \ldots, N$, onde $h = \frac{L}{N}$ e, dividindo o intervalo $[0, T]$, da variável tempo t, em M partes iguais de comprimento k, temos os $M + 1$ pontos $t_j = jk, j = 0, 1, \ldots, M$, onde $k = \frac{T}{M}$. Assim vamos obter uma aproximação da solução nos pontos (x_i, t_j) da malha, como mostra a Figura 13.2:

Figura 13.2: Domínio da equação e respectiva malha.

Vamos denotar por $u_{i,j}$ a solução exata no ponto (x_i, t_j) e por $U_{i,j}$ um valor aproximado de $u_{i,j}$.

13.3 Métodos de Diferenças Finitas

Método Explícito

Usando diferenças centradas de segunda ordem na variável espacial para aproximar a derivada de segunda ordem obtemos:

$$u_{xx}(x_i, t_j) \simeq \frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h^2} = \delta_x^2 U_{i,j} \ . \tag{13.15}$$

e usando agora diferenças progressivas no tempo para aproximar a derivada de primeira ordem produzimos a aproximação:

$$u_t(x_i, t_j) \simeq \frac{U_{i,j+1} - U_{i,j}}{k} = \Delta_t U_{i,j} \ .$$

Substituindo essas aproximações em (13.2), obtemos a equação aproximada:

$$\frac{U_{i,j+1} - U_{i,j}}{k} = \alpha \left(\frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h^2} \right) \ , \tag{13.16}$$

ou seja,

$$U_{i,j+1} = U_{i,j} + \sigma(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}), \tag{13.17}$$

onde $\sigma = k\alpha/h^2$.

Assim, conhecidos os valores $U_{i-1,j}, U_{i,j}$ e $U_{i+1,j}$ calculamos $U_{i,j+1}$ explicitamente, sem qualquer complicação suplementar (ver Figura 13.3):

Figura 13.3: Discretização e correspondente molécula computacional do método explícito.

A molécula computacional é uma tradução gráfica da fórmula (13.17), pois ela estabelece a relação de dependência existente entre o valor no ponto $(i, j+1)$ e seus vizinhos. Note que no caso do método explícito o ponto $(i, j+1)$ depende apenas dos pontos $(i-1, j)$, (i, j) e $(i+1, j)$ todos do nível anterior e daí a palavra explícito. Observe também na Figura 13.3 que, como os valores sobre a linha $j = 0$ são conhecidos (dados iniciais), é possível calcular os valores da linha $j = 1$ a menos do primeiro e do último, mas esses dois valores são dados exatamente pelas condições de fronteira completando assim o cálculo da linha $j = 1$. Tendo a linha $j = 1$ procedemos de maneira análoga para calcular a linha $j = 2, \ldots$. Ilustraremos essas idéias através de um exemplo.

Exemplo 13.1 - *Calcule a primeira linha de soluções da equação a seguir , com* $\sigma = \frac{1}{6}$ *e* $k = \frac{1}{54}$.

$$\begin{aligned} u_t &= u_{xx} \ , & 0 \le x \le 1 \ , \ \ 0 < t < T \ , \\ u(x,0) &= x(1-x) \ , & 0 \le x \le 1 \ , \\ u(0,t) &= 0 \ , & 0 < t < T \ , \\ u(1,t) &= 0 \ , & 0 < t < T \ , \end{aligned} \tag{13.18}$$

cuja solução exata é:

$$u(x,t) \ = \ \frac{8}{\pi^3} \sum_{n}^{\infty} \frac{sen\,(2n+1)\pi x}{(2n+1)^3} \exp(-((2n+1)\pi)^3 t) \ . \tag{13.19}$$

Solução: Como $\sigma = \frac{1}{6}$, $\alpha = 1$ e $k = \frac{1}{54}$ temos que $h = \frac{1}{3}$ e portanto $x_0 = 0$, $x_1 = \frac{1}{3}$, $x_2 = \frac{2}{3}$, $x_3 = 1$ e $t_0 = 0$, $t_1 = \frac{1}{54}, t_2 = \frac{2}{54}, \ldots$. Da condição inicial vem que:

$$\begin{aligned}
u_{00} &= u(x_0,t_0) = u(0,0) = 0(1-0) = 0 = U_{00} \ , \\
u_{10} &= u(x_1,t_0) = u(\frac{1}{3},0) = \frac{1}{3}(1-\frac{1}{3}) = \frac{2}{9} = U_{10} \ , \\
u_{20} &= u(x_2,t_0) = u(\frac{2}{3},0) = \frac{2}{3}(1-\frac{2}{3}) = \frac{2}{9} = U_{20} \ , \\
u_{30} &= u(x_3,t_0) = u(1,0) = 1(1-1) = 0 = U_{30} \ .
\end{aligned}$$

Das condições de fronteira deduz-se que:

$$\begin{aligned}
u_{01} &= u(x_0,t_1) = u(0,\frac{1}{54}) = 0 = U_{01} \ , \\
u_{31} &= u(x_3,t_1) = u(1,\frac{1}{54}) = 0 = U_{31} \ .
\end{aligned}$$

Usando o método explícito:

$$\begin{aligned}
u_{11} &= u(x_1,t_1) = 0.1861023 \simeq U_{11} = U_{10} + \sigma(U_{00} - 2U_{10} + U_{20}) = \frac{5}{27} = 0.1851852 \ , \\
u_{21} &= u(x_2,t_1) = 0.1861023 \simeq U_{21} = U_{20} + \sigma(U_{10} - 2U_{20} + U_{30}) = \frac{5}{27} = 0.1851852 \ .
\end{aligned}$$

Erro de Truncamento Local

O exemplo (13.3) acima mostra claramente que a solução numérica calculada nos pontos (x_1,t_1) e (x_2,t_1) são apenas aproximações para o valor verdadeiro da solução nesses pontos, apesar de termos utilizado a solução exata no cálculo dos valores anteriores que entram na formação de $U_{1,1}$ e $U_{2,1}$. Dessa forma, o erro introduzido no cálculo acima advém única e exclusivamente da substituição das derivadas por diferenças finitas, ou em última instância da substituição da equação diferencial pela equação de diferenças. A esse erro chamaremos de **Erro de Truncamento Local**, que passaremos a definir precisamente.

Denotando por $u_{i,j} = u(x_i,t_j)$ e por $\tau_{i,j}$ o erro ocorrido no cálculo de $U_{i,j+1}$ assumindo que todos os valores anteriores utilizados nesse cálculo são exatos, e ponderado pelo passo temporal k, podemos definir:

$$\tau_{i,j} = \frac{u(x_i,t_{j+1}) - U_{i,j+1}}{k} = \frac{u(x_i,t_{j+1}) - (U_{i,j} + \sigma(U_{i-1,j} - 2Ui,j + U_{i+1,j}))}{k} \ ,$$

onde utilizamos a equação de diferenças (13.17) para substituir o valor de $U_{i,j+1}$. Usando agora a hipótese de que $U_{i,j} = u(x_i,t_j)$, $\forall i$ temos:

$$\tau_{i,j} = \frac{u(x_i,t_{j+1}) - (u(x_i,t_j) + \sigma(u(x_{i-1},t_j) - 2u(x_i,t_j) + u(x_{i+1},t_j)))}{k} \ ,$$

que, substituindo o valor de σ pode ser reescrita na forma:

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{\alpha}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + \tau_{i,j} \ . \qquad (13.20)$$

Observe que (13.20) tem exatamente a mesma forma de (13.16) a menos do termo do erro de truncamento local, o que nos permite dizer que o bf Erro de Truncamento Local é uma medida de quanto a solução da equação diferencial, discretizada na malha, deixa de satisfazer a equação de diferenças. Note que a situação inversa, isto é, quanto a solução da equação de diferenças deixa de satisfazer a diferencial não é possível de ser definida, uma vez que a primeira sendo uma solução discreta não pode ser diferenciada para substituição na equação diferencial. Portanto o inverso do erro de truncamento local não pode ser definido.

Uma forma mais prática para o erro de truncamento local pode ser obtida aplicando expansão em série de Taylor em torno de (x_i, y_j).

$$
\begin{aligned}
u_{i,j+1} &= u(x_0+ih, t_0+jk+k) \\
&= u(x_0+ih, t_0+jk) + k\frac{\partial u}{\partial t}(x_0+ih, t_0+jk) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_0+ih, t_0+jk) + O(k^3) \ , \\
u_{i+1,j} &= u(x_0+ih+h, t_0+jk) \\
&= u(x_0+ih, t_0+jk) + h\frac{\partial u}{\partial x}(x_0+ih, t_0+jk) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_0+ih, t_0+jk) \\
&+ \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x_0+ih, t_0+jk) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(x_0+ih, t_0+jk) + O(k^5) \ , \\
u_{i-1,j} &= u(x_0+ih-h, t_0+jk) \\
&= u(x_0+ih, t_0+jk) - h\frac{\partial u}{\partial x}(x_0+ih, t_0+jk) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_0+ih, t_0+jk) \\
&- \frac{h^3}{3!}\frac{\partial^3 u}{\partial x^3}(x_0+ih, t_0+jk) + \frac{h^4}{4!}\frac{\partial^4 u}{\partial x^4}(x_0+ih, t_0+jk) + O(k^5) \ .
\end{aligned}
$$

Substituindo na equação (13.20), e cancelando os termos comuns obtemos:

$$
u_t + \frac{k}{2}u_{tt} + O(k^2) \ = \ \alpha u_{xx} + \alpha\frac{h^2}{12}u_{xxxx} + O(h^3) + \tau_{i,j} \ .
$$

Daí utilizando a equação diferencial $u_t = \alpha u_{xx}$, podemos reescrever o erro de truncamento local como:

$$
\tau_{i,j} \ = \ \frac{k}{2}u_{tt} - \frac{\alpha h^2}{12}u_{xxxx} + O(k^2) + O(h^3) \ = \ O(k+h^2) \ . \tag{13.21}
$$

Erro Global, Estabilidade, Consistência e Convergência

Definimos o *erro global* em um ponto (x_i, t_j), por:

$$
e_{i,j} \ = \ U_{i,j} - u_{i,j} \ ,
$$

isto é, a diferença entre a solução aproximada e a exata no ponto (x_i, t_j) da malha. Note que diferentemente do caso do erro de truncamento local a definição de erro global não assume que os valores anteriores utilizados no cálculo de $U_{i,j}$ são exatos, e portanto o erro global como o nome sugere, pode conter toda espécie de erro que contamine a solução incluindo o erro de arredondamento do computador.

Então, da equação (13.17), temos:

$$\begin{aligned} e_{i,j+1} &= U_{i,j+1} - u_{i,j+1} \\ &= U_{i,j} + \sigma(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}) - [u_{i,j} + \sigma(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + k\tau_{i,j}] \\ &= e_{i,j} + \sigma(e_{i-1,j} - 2e_{i,j} + e_{i+1,j}) - k\tau_{i,j} \\ &= \sigma(e_{i-1,j} + e_{i+1,j}) + (1-2\sigma)e_{i,j} - k\tau_{i,j} \ , \end{aligned} \tag{13.22}$$

e, então

$$|e_{i,j+1}| \ \leq \ |\sigma|(|e_{i-1,j}| + |e_{i+1,j}|) + |1-2\sigma||e_{i,j}| + |-k||\tau_{i,j}| \ . \tag{13.23}$$

Fazendo $E_j = \max\{|e_{p,j}|, 0 \leq p \leq N\}$, $\tau_j = \max\{|\tau_{p,j}|, 0 \leq p \leq N\}$, supondo $1 - 2\sigma \geq 0$, (condição de estabilidade) e como $\sigma > 0$ podemos reescrever a equação (13.23) como:

$$|e_{i,j+1}| \ \leq \ |\sigma| 2E_j + |1-2\sigma|E_j + |k||\tau_{i,j}| \leq E_j + k \mid \tau_{i,j} \mid \leq E_j + \tau_j \ .$$

Portanto:

$$E_{j+1} \ \leq \ E_j + k\tau_j \ . \tag{13.24}$$

Aplicando (13.24) recursivamente para $j, j-1, j-2, \ldots, 1$, obtemos a seguinte expressão:

$$E_{j+1} \ \leq \ k(\tau_0 + \tau_1 + \ldots + \tau_j) \leq (j+1)k\tau \leq T\tau \ ,$$

onde $\tau = \max\{\tau_i, i = 0, 1, \ldots, M\}$ e $T = Mk$ é um limitante para o domínio na direção do eixo tempo, ver equação (13.14).

Então, se a condição $1 - 2\sigma \geq 0$ é satisfeita e se τ é de ordem pelo menos h, $E_{j+1} \to 0$ quando $k \to 0$.

Observe que para provarmos que $E_j \to 0$ foi necessário assumir duas condições: $\tau = O(h)$ e $1-2\sigma \geq 0$. Essas duas hipóteses são cruciais para a conclusão do resultado, sem elas ele não pode ser provado. Essas hipóteses são na verdade os conceitos de **Consistência** e **Estabilidade** que passamos a definir mais precisamente.

Definição 13.1 - *Um método numérico é* **consistente** *com relação a uma dada equação se o erro de truncamento local desse método para aquela equação for pelo menos de* $O(h)$

Definição 13.2 - *Um método numérico é* **estável** *se a equação de diferenças associada não amplifica erros dos passos anteriores.*

Por exemplo a equação de diferenças $y_{j+1} = \sigma y_j$ é est ável se $|\sigma| \leq 1$ e instável se $|\sigma| > 1$ pois: Sejam y_j e z_j as soluções dessa equação com os dados iniciais y_0 e $z_0 = y_0 + \epsilon$, onde ϵ é um número pequeno. O leitor não terá dificuldade em mostrar que

$$y_j = (\sigma)^j y_0, \quad z_j = (\sigma)^j z_0 = (\sigma)^j (y_0 + \epsilon) = (\sigma)^j y_0 + (\sigma)^j \epsilon \ .$$

E portanto,

$$|z_j - y_j| = (\sigma)^j \epsilon$$

,

ou seja o erro ϵ cometido no primeiro passo é amortecido ou amplificado dependendo de $|\sigma|$.

Definição 13.3 - *Um método numérico é* **convergente** *num ponto* (x,t) *do domínio se o erro global* $E_{i,j}$ *associado com esse ponto tende a zero quando os indices* i *e* j *tendem para infinito de maneira que o ponto* $x = i*h$ *e* $t = j*k$ *permaneça fixo.*

O teorema de equivalência de LAX estabelece que para equações lineares as propriedades de estabilidade e consistência são equivalentes àquela de convergência. (citar referencia)

Voltando ao exemplo do método explícito observamos que desigualdade $1-2\sigma \geq 0$, pode ser reescrita como $\sigma \leq \frac{1}{2}$, que é a condição para estabilidade para esse método, que será então chamado de **condicionalmente estável**.

Chamamos atenção também para o fato de que a estabilidade é uma propriedade intrinseca da equação de diferenças finitas, e consiste em que esta não tenha o defeito de amplificar erros dos passos anteriores. Já com relação aos critérios para determinação da estabilidade, estudaremos a seguir os dois mais conhecidos: o de Von Neumann e o da matriz.

Estabilidade - Critério de von Neumann

Este é um critério simples e muito utilizado para determinar a estabilidade de um método numérico. Ele é baseado no princípio da superposição, ou seja na observação de que o erro global é a somatória de erros mais simples também chamados harmônicos. Esse processo é inspirado na expansão de uma função em série de Fourier. Denotando por $E_i, \; i = 0, 1, \ldots N$ o erro global em cada ponto ao longo da primeira linha $t = 0$ podemos escrever:

$$E_i = \sum_{n=0}^{N} a_n \exp(I\alpha_n ih), \quad i = 0, 1, \ldots N \; ,$$

onde $I = \sqrt{-1}, \;\; \alpha_n = \frac{n\pi}{L}$ e $Nh = L$, o que constitui um sistema linear com $N+1$ incógnitas a_n e $N+1$ equações, cuja matriz dos coeficientes é não singular, e portanto pode ser resolvido de maneira única para determinar a_n. Tendo representado o erro no passo inicial, para analisar sua propagação ao longo dos passos subsequentes basta observar a propagação de um harmônico genérico $\exp(I\beta ih)\exp(\lambda jk)$ onde β é um número real e λ um número complexo, ambos arbitrários.

Portanto a estratégia do critério de von Neumann para determinação da estabilidade é a de examinar o resultado da propagação de um dado modo de Fourier em uma linha ou estágio subsequente j. Se houve amplificação desse harmônico dizemos que o método é instável, se houve amortecimento ele será estável.

Por utilizar o princípio de superposição o critério de von Neumann só deve ser usado quando a equação é linear com coeficientes constantes e além disso ele ignora completamente a influência das condições de fronteira sobre o comportamento da solução da equação de diferenças. Geralmente, a condição de estabilidade deduzida do critério de von Neumann produz uma condição necessária para a estabilidade, mas não suficiente, uma discussão bastante detalhada desse problema é apresentada em [**?**] páginas 117-132, veja também o exercício (13.22).

A seguir ilustramos a aplicação prática do critério de von Neumann utilizando-o para determinar a estabilidade do método explícito. Com esse objetivo vamos admitir então que exista uma solução da equação de diferenças (13.17) da forma:

$$U_{i,j} \;=\; e^{\lambda j} e^{I\beta i} = (e^{\lambda})^j e^{I\beta i} \; , \tag{13.25}$$

e tentamos encontrar λ e β tais que (13.25) seja de fato uma solução de (13.17). Substituimos (13.25) em (13.17) para obter:

$$e^{\lambda(j+1)} e^{I\beta i} = (1-2\sigma)U_{i,j} + \sigma(e^{\lambda j} e^{I\beta(i-1)} + e^{\lambda j} e^{I\beta(i+1)}) \; .$$

Assim:

$$e^{\lambda} U_{i,j} \;=\; (1-2\sigma)U_{i,j} + \sigma(e^{-I\beta}U_{i,j} + e^{I\beta}U_{i,j}) \;.$$

Logo, eliminando os termos comuns obtem-se:

$$\begin{aligned} e^{\lambda} &= (1-2\sigma) + \sigma(e^{-I\beta} + e^{I\beta}) \\ &= (1-2\sigma) + 2\sigma\cos\beta \\ &= 1 + 2\sigma(\cos\beta - 1) \\ &= 1 - 4\sigma \ \text{sen}^2\frac{\beta}{2} \;. \end{aligned}$$

Como $\sigma \geq 0$ então $e^{\lambda} = 1 - 4\sigma \ \text{sen}^2\frac{\beta}{2} \leq 1$. Assim, se $e^{\lambda} \geq 0$, de (13.25) a solução da equação (13.17) decairá uniformente quando $j \to \infty$. No entanto e^{λ} pode ser negativo, uma vez que λ é complexo, e portanto teremos mais duas situações a considerar: $-1 \leq e^{\lambda} < 0$ a solução terá amplitude decrescente e sinal oscilante quando $j \to \infty$. Finalmente, se $e^{\lambda} < -1$ a solução oscila com amplitude crescente quando $j \to \infty$. Neste último caso (13.17) é instável, enquanto que no caso anterior ela será estável. Assim, resumindo, para estabilidade será exigido que:

$$|e^{\lambda}| \;\leq\; 1 \;.$$

Como $e^{\lambda} < 1$ sempre, precisamos ainda impor $-1 \leq e^{\lambda}$, ou seja, $-1 \leq 1 - 4\sigma \ \text{sen}^2\frac{\beta}{2}$.

Portanto:

$$\sigma \;\leq\; \frac{1}{1-\cos\beta} \quad \forall\beta, \;\; \text{ou seja} \;\; \sigma \leq \frac{1}{2} \;. \tag{13.26}$$

Note que a condição (13.26) obtida pelo critério de von Neumann para estabilidade do método explícito é exatamente aquela que obtemos na seção anterior impondo que $1 - 2\sigma \geq 0$.

Com a imposição do limitante sobre σ para estabilidade, o método explícito geralmente produz aproximações satisfatórias. Porém, $\sigma < \frac{1}{2}$ é uma condição muito restritiva para o tamanho do passo na direção t, pois esta condição significa que $k < \frac{h^2}{2\alpha}$, e o esforço computacional poderá ser grande se desejarmos calcular a solução para um tempo T razoavelmente longo.

Estabilidade - Critério da Matriz

Vamos iniciar esta seção observando que a discretização explícita da equação (13.14) e respectivas condições iniciais e de fronteira fornece a seguinte equação de diferenças:

$$\begin{aligned} U_{i,j+1} &= \sigma U_{i-1,j} + (1-2\sigma)U_{i,j} + \sigma U_{i-1,j}, \quad i = 1,2,\ldots N-1, \quad j = 0,1,\ldots \;, \\ U_{i,0} &= \psi(ih), \quad i = 0,1,\ldots N \;, \\ U_{0,j} &= f(jk), \quad j = 1,2,\ldots \;, \\ U_{N,j} &= g(jk), \quad j = 1,2,\ldots \;. \end{aligned} \tag{13.27}$$

Introduzindo a notação vetorial:

$$\mathbf{U}_j \;=\; (U_{1,j}, U_{2,j}, \cdots U_{N-1,j})^T \;, \tag{13.28}$$

então para cada j a equação (13.27) pode ser escrita na forma matricial:

$$\mathbf{U}_{j+1} \;=\; A\mathbf{U}_j + \mathbf{c}_j, \quad j = 0, 1, \ldots \;, \tag{13.29}$$

onde A é a seguinte matriz $(N-1) \times (N-1)$:

$$A \;=\; \begin{pmatrix} 1-2\sigma & \sigma & 0 & \ldots & 0 \\ \sigma & 1-2\sigma & \sigma & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & \sigma & 1-2\sigma & \sigma \\ 0 & \ldots & 0 & \sigma & 1-2\sigma \end{pmatrix} , \tag{13.30}$$

e $\mathbf{c}_j$ é um vetor de $(N-1)$ componentes contendo informações das condições de fronteira dado por:

$$\mathbf{c}_j \;=\; (f(jk), 0, \ldots, 0, g(jk))^T \;.$$

De maneira análoga introduzindo os vetores $\mathbf{u}_j = (u(x_1, t_j), \ldots, u(x_{N-1}, t_j))^T$, $\tau_j = (\tau_{1,j}, \ldots, \tau_{N-1,j})^T$ e $\mathbf{e}_j = (e_{1,j}, \ldots, e_{N-1,j})^T$ para representarem, respectivamente, a solução exata, o erro de truncamento local e o erro global, e considerando (13.20), podemos escrever a equação matricial para o erro de truncamento local:

$$\mathbf{u}_{j+1} \;=\; A\mathbf{u}_j + \mathbf{c}_j + \tau_j \;. \tag{13.31}$$

Subtraindo (13.29) de (13.31) obtemos a equação vetorial para o erro global:

$$\mathbf{e}_{j+1} \;=\; A\mathbf{e}_j + \tau_j \;. \tag{13.32}$$

Aplicando a equação (13.32) recursivamente para $j,\ j-1,\ \ldots 0$ obtemos:

$$\begin{aligned} \mathbf{e}_{j+1} &= A^{j+1}\mathbf{e}_0 + A^j\tau_0 + A^{j-1}\tau_1 + \cdots + A\tau_{j-1} + \tau_j \\ &= A^j\tau_0 + A^{j-1}\tau_1 + \cdots + A\tau_{j-1} + \tau_j \;, \end{aligned} \tag{13.33}$$

se lembrarmos que $\mathbf{e}_0 = 0$ por construção. Vê-se de (13.34) que o erro global é formado pelo acúmulo dos erros de truncamento local de cada passo propagados pelas potências da matriz A. Portanto a matriz A tem um papel crucial na propagação desses erros e ela é chamada de *matriz de amplificação.* O erro cresce se algum autovalor de A tem módulo maior do que 1. Se todos são menores do que 1 em módulo, temos o erro decrescendo e portanto estabilidade. Definimos então:

Definição 13.4 - *Uma equação vetorial de diferenças da forma:*

$$\mathbf{U}_{j+1} \;=\; A\mathbf{U}_j + \mathbf{c}_j \;,$$

é **estável** *com relação a alguma norma* $||\cdot||$ *se e sómente se existem constantes* $h_0 > 0$*,* $k_0 > 0$*,* $K \geq 0$ *e* $\beta \geq 0$ *com:*

$$||A^n|| \;\leq\; K\exp(\beta t) \;,$$

sempre que $0 \leq t = nk,\ 0 < h \leq h_0$ *e* $0 < k \leq k_0$*. Em particular se os autovetores de A são linearmente independentes e se seus autovalores* λ_i *satisfazem* $|\lambda_i| \leq 1 + O(k)$*,* $\forall i$*, então o método será estável. Ver exercício (13.10).*

No caso particular da matriz A do método explícito (13.30) seus autovalores são dados no exercício (13.12) e portanto para estabilidade precisamos que:

$$|\lambda_i| \; = \; |1 - 4\sigma \, \text{sen}^2(\frac{i\pi}{2N})| \leq 1 + O(k) \; ,$$

que pode ser facilmente mostrado implicar em $\sigma \leq \frac{1}{2}$, ou seja a mesma condição já obtida pelos critérios anteriores.

Note que a matriz A de um método numérico geral terá sempre seus elementos dependentes do parâmetro σ, e portanto determinar a estabilidade desse método numérico requer a determinação dos autovalores de uma matriz de ordem arbitrária N cujos elementos dependem de σ. Esta pode ser uma tarefa bastante difícil. Para tanto contamos com alguns resultados de álgebra linear que nos auxiliam na tarefa de encontrar limitantes para esses autovalores. Eles são os teoremas de Gerschgorin que aplicam-se para uma matriz A geral da forma:

$$A \; = \; \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \ldots & a_{NN} \end{pmatrix} .$$

Seja $d_i = \sum_{j=1}^{N} |a_{i,j}|$, temos então o primeiro teorema:

Teorema 13.2 *Seja λ o maior autovalor de A tomado em módulo. Então $\lambda \leq \max_{1 \leq i \leq N} |d_i|$.*

Prova: Ver [**?**] página 87.

O segundo teorema pode ser enunciado como a seguir:

Teorema 13.3 *Seja $r_i = d_i - |a_{ii}|$. Então os autovalores de A estão na união dos discos de centro em $|a_{ii}|$ e raio r_i.*
*Prova: Ver [**?**] páginas 88-89.*

Como mostrado em (13.21) o erro de truncamento local do método explícito é de ordem $O(k + h^2)$. Uma pergunta que surge naturalmente é se podemos obter métodos com ordem melhor. A resposta é positiva como veremos a seguir.

Melhorando a ordem

Uma tentativa de melhoria da ordem, ainda mantendo a condição de método explícito, pode ser feita incorporando-se mais um ponto à fórmula (13.17), ou seja, usando-se diferença centrada na variável t. Tem-se então a seguinte equação de diferenças, chamado **Método de Richardson**, e respectiva molécula computacional:

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} \; = \; \alpha \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} \; . \tag{13.34}$$

e o erro de truncamento local será:

$$\tau_{i,j} \;=\; \frac{k^2}{6} u_{ttt}(x_i, \eta) - \frac{h^2}{12} u_{xxxx}(\xi, t_j) = O(h^2 + k^2) \; ,$$

portanto de ordem 2.

A análise da estabilidade pode ser feita utilizando o critério de von Neumann. Escrevemos então:

$$U_{i,j} \;=\; e^{\lambda j} e^{I\beta i} \; ,$$

que após substituição em (13.34) e várias simplificações resulta em:

$$\lambda \;=\; -4\sigma \operatorname{sen}^2 \frac{\beta}{2} \pm (1 + 16\sigma^2 \operatorname{sen}^4 \frac{\beta}{2})^{\frac{1}{2}} \; .$$

Teremos sempre uma raiz de módulo maior do que 1, portanto este método é incondicionalmente instável.

Uma opção para solucionar o problema da instabilidade é substituir o termo $U_{i,j}$ pela média dos termos $U_{i,j+1}$ e $U_{i,j-1}$, na aproximação de u_{xx}. Vamos obter dessa forma o **Método de Du Fort-Frankel**:

$$\frac{U_{i,j+1} - U_{i,j-1}}{2k} = \alpha \frac{U_{i+1,j} - (U_{i,j+1} + U_{i,j-1}) + U_{i-1,j}}{h^2} \; ,$$

cujo erro de truncamento local é dado por:

$$\tau_{i,j} = \frac{k^2}{6} u_{ttt} - \frac{h^2}{12} u_{xxxx} - \frac{k^2}{h^2} \alpha u_{tt} + O(h^4 + k^4 + \frac{k^4}{h^2}) \; .$$

Agora, se $k = rh$ temos que o método *não* será consistente, melhor dizendo, o método será consistente com a equação:

$$u_t = \alpha u_{xx} - \alpha r^2 u_{tt} \; ,$$

que é uma equação hiperbólica!

Portanto, para obtermos um método consistente, é preciso restringir k em função de h, por exemplo $k = rh^2$ e aí teremos um método consistente de ordem 2. O método de Du Fort-Frankel é pois condicionalmente consistente e a condição imposta é bastante restritiva.

Analisando a estabilidade, vemos que, a partir de

$$U_{i,j+1} - U_{i,j-1} = 2\sigma(U_{i+1,j} - (U_{i,j+1} + U_{i,j-1}) + U_{i-1,j}) \; ,$$

obtem-se:

$$(1 + 2\sigma)U_{i,j+1} - 2\sigma(U_{i+1,j} + U_{i-1,j}) - (1 - 2\sigma)U_{i,j-1} = 0 \; ,$$

e pelo critério de von Neumann, escrevendo:

$$U_{i,j} = e^{\lambda j} e^{I\beta i} \ ,$$

teremos:

$$(1 + 2\sigma)e^{2\lambda} - 2e^{\lambda}\sigma \cos\beta - (1 - 2\sigma) = 0 \ ,$$

cujas raízes são:

$$\gamma = \frac{2\sigma \cos\beta \pm (1 - 4\sigma \operatorname{sen}^2\beta)^{\frac{1}{2}}}{1 + 2\sigma} = \frac{a \pm Ib}{1 + 2\sigma} \ .$$

Temos que $1 - 4\sigma \operatorname{sen}^2\beta < 1$, podendo ser positivo ou negativo.

i) Se negativo: Teremos raízes complexas conjugadas e

$$|\gamma| = \frac{(a + Ib)(a - Ib)}{1 + 2\sigma} = \frac{4\sigma^2 - 1}{4\sigma^2 + 4\sigma + 1} < 1 \ .$$

ii) Se positivo: Então,

$$0 < (1 - 4\sigma \operatorname{sen}^2\beta)^{\frac{1}{2}} < 1 \ \text{ e } \ |2\sigma \cos\beta| < 2\sigma \ , \ \Rightarrow |\gamma| < \frac{1 + 2\sigma}{2\sigma + 1} = 1.$$

Portanto o método é incondicionalmente estável.

Este método é um método explícito de dois passos, pois exige conhecimento da solução em 2 níveis anteriores de tempo para a construção da solução no próximo nível. Diferentemente dos métodos de um nível, ele só pode ser aplicado para o cálculo da solução a partir do nível 2, exigindo que o primeiro passo seja calculado por meio de outro método.

Este método constitui um exemplo claro do cuidado que se deve ter em relação ao teorema da equivalência de Lax ilustrando o importante fato de que nem sempre a escolha da malha (relação entre h e k) é ditada pela condição de estabilidade. No caso do método de Dufort Frankel quem impõe a escolha da malha é a consistência e não a estabilidade. Está implícito no teorema de Lax que é preciso verificar não só a condição de estabilidade mas também a escolha adequada da malha.

Como vimos, resolvemos o problema da instabilidade e criamos um outro com a consistência, este será sempre um dilema pois essas duas propriedades estão na maioria das vezes em conflito. Uma maneira de aliviar esse problema é considerar métodos implícitos, pois estes têm melhores propriedades de estabilidade, mas como veremos abaixo o custo computacional aumenta consideravelmente.

Método Implícito

Uma **fórmula implícita** é aquela em que dois ou mais valores desconhecidos na linha j são especificados em termos de valores conhecidos das linhas $j-1, j-2, \ldots$. Claramente, não será possível o cálculo direto de $U_{i,j}$; usualmente exige-se a solução de um sistema linear.

Aplicando diferenças regressivas do lado esquerdo da equação e diferenças centradas do lado direito, obtemos:

$$\frac{U_{i,j} - U_{i,j-1}}{k} = \alpha \left(\frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h^2} \right) \ , \tag{13.35}$$

que depois de alguma manipulação algébrica pode ser reescrita na forma:

$$-\sigma U_{i-1,j} + (1 + 2\sigma)U_{i,j} - \sigma U_{i+1,j} = U_{i,j-1} \ \text{ para } \ i = 1, 2, \ldots, N-1 \ \text{ e } \ j = 1, 2, \ldots \ . \tag{13.36}$$

A molécula computacional dá-nos uma idéia precisa da relação de dependência entre os diversos elementos da fórmula.

Figura 13.5: Molécula computacional do método implícito.

Observe que (13.36) forma um sistema tridiagonal de equações, e ao resolvê-lo encontramos todas as aproximações do estágio j. Escrevendo mais detalhadamente, o sistema a ser resolvido em cada estágio é:

$$\begin{pmatrix} 1+2\sigma & -\sigma & 0 & \dots & 0 \\ -\sigma & 1+2\sigma & -\sigma & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & -\sigma & 1+2\sigma & -\sigma \\ 0 & \dots & 0 & -\sigma & 1+2\sigma \end{pmatrix} \begin{pmatrix} U_{1j} \\ U_{2j} \\ \vdots \\ U_{N-2,j} \\ U_{N-1,j} \end{pmatrix} = \begin{pmatrix} U_{1,j-1} \\ U_{2,j-1} \\ \vdots \\ U_{N-2,j-1} \\ U_{N-1,j-1} \end{pmatrix} + \begin{pmatrix} \sigma U_{0,j} \\ 0 \\ \vdots \\ 0 \\ \sigma U_{N,j} \end{pmatrix} ,$$

ou na notação vetorial introduzida em (13.28):

$$A\mathbf{U}_j = \mathbf{U}_{j-1} + \mathbf{c}_j \ , \qquad (13.37)$$

onde $\mathbf{c}_j = (\sigma f(jh), 0, \dots, 0, \sigma g(jh))^T$, para o caso de condições de fronteira como as de (13.14).

Podemos observar que A é uma matriz $(N-1) \times (N-1)$ estritamente diagonalmente dominante e portanto o sistema (13.37) tem solução única. Além disso, a propriedade de diagonal dominância nos garante que (13.37) pode ser resolvido por qualquer método de solução de sistemas de equações lineares, como os métodos dados nos Capítulos 4 e 5.

Exemplo 13.2 - *Calcule a primeira linha de soluções do problema do exercício (13.19)*

Solução: Temos:

$$\sigma = \frac{1}{6} \text{ e } k = \frac{1}{54} \Rightarrow h = \frac{1}{3} \ .$$

$$\text{Com } h = \frac{1}{3} \text{ temos } x_0 = 0, x_1 = \frac{1}{3}, x_2 = \frac{2}{3}, x_3 = 1 \ .$$

$$\text{Com } k = \frac{1}{54} \text{ temos } t_0 = 0, t_1 = \frac{1}{54}, t_2 = \frac{2}{54}, \dots \ .$$

Usando as aproximações $U_{00}, U_{10}, U_{20}, U_{30}, U_{01}$ e U_{31}, do exemplo (13.19), vamos calcular U_{11} e U_{21}.

$$U_{10} = -\sigma U_{01} + (1+2\sigma)U_{11} - \sigma U_{21} \ ,$$
$$U_{20} = -\sigma U_{11} + (1+2\sigma)U_{21} - \sigma U_{31} \ .$$

$$\begin{pmatrix} (1+2\sigma) & -\sigma \\ -\sigma & (1+2\sigma) \end{pmatrix} \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} = \begin{pmatrix} U_{10} + \sigma U_{01} \\ U_{20} + \sigma U_{31} \end{pmatrix} ,$$

$$\begin{pmatrix} 4/3 & -1/6 \\ -1/6 & 4/3 \end{pmatrix} \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} = \begin{pmatrix} 2/9 \\ 2/9 \end{pmatrix} .$$

Resolvendo o sistema acima obtemos:

$$U_{11} = U_{21} = 0.1904762.$$

A solução exata é $u_{11} = u_{21} = 0.1861023$.

Veja que a utilização de um método implícito como o acima é de difícil justificativa prática, uma vez que a precisão dos resultados não melhorou na mesma proporção do aumento do esforço computacional. Apresentaremos ainda neste capítulo um método implícito que apresenta erro de truncamento com ordem mais alta. Antes vamos calcular o erro de truncamento do método implícito para justificar os resultados numéricos obtidos no exemplo acima.

Erro de Truncamento Local

O desenvolvimento aqui segue exatamente as mesmas idéias do caso do método explícito e portanto repetiremos sómente os passos principais.

Seja $\tau_{i,j}$ o erro de truncamento ocorrido no cálculo de $U_{i,j}$. Então, podemos escrever:

$$\frac{u_{i,j} - u_{i,j-1}}{k} = \frac{\alpha}{h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) + \tau_{i,j} \ .$$

Aplicando expansão em série de Taylor em torno de (x_i, y_j), obtemos,

$$u_t - \frac{k}{2}u_{tt} + O(k^2) = \alpha u_{xx} + \alpha\frac{h^2}{12}u_{xxxx} + O(h^3) + \tau_{i,j} \ .$$

Assim:

$$\tau_{i,j} = -\frac{k}{2}u_{tt} - \frac{\alpha h^2}{12}u_{xxxx} + O(k^2) + O(h^3) = O(k + h^2) \ . \tag{13.38}$$

Concluímos que o método é **incondicionalmente consistente** e de ordem 1.

Comparando as expressões (13.21) e (13.38) observamos que elas diferem apenas no sinal do termo $\frac{k}{2}u_{tt}$ sendo um positivo e outro negativo. Esse fato nos motiva a considerar uma nova aproximação que é a média entre as aproximações explícita e implícita, na esperança de que esse termo despareça do erro de truncamento local, como de fato ocorre. Essa estratégia dá origem ao método de Crank-Nicolson que será estudado logo mais adiante.

Erro Global

Segue o mesmo desenvolvimento feito para o método explícito.

Estabilidade

Analogamente ao caso do método explícito pelo critério de von Neumann escrevemos:

$$U_{i,j} = e^{\lambda j} e^{I\beta i} \ .$$

Substituindo na equação (13.36) temos o seguinte desenvolvimento:

$$\begin{aligned}
U_{i,j-1} &= -\sigma U_{i-1,j} + (1+2\sigma)U_{i,j} - \sigma U_{i+1,j} \\
e^{\lambda(j-1)} e^{I\beta i} &= -\sigma e^{\lambda j} e^{I\beta(i-1)} + (1+2\sigma)U_{i,j} - \sigma e^{\lambda j} e^{I\beta(i+1)} \\
e^{-\lambda} U_{i,j} &= -\sigma e^{-I\beta} U_{i,j} + (1+2\sigma)U_{i,j} - \sigma e^{I\beta} U_{i,j} \\
e^{-\lambda} &= -\sigma(e^{-I\beta} + e^{I\beta}) + (1+2\sigma) \\
&= -2\sigma\cos\beta + 1 + 2\sigma \\
&= 1 + 2\sigma(1-\cos\beta) \\
&= 1 + 4\sigma \,\mathrm{sen}^{\,2}\frac{\beta}{2}.
\end{aligned}$$

Portanto:

$$e^{\lambda} = \frac{1}{1 + 4\sigma\,\mathrm{sen}^{\,2}\frac{\beta}{2}} \ .$$

Para a estabilidade, $|e^{\lambda}| \leq 1$. Mas, neste caso $|e^{\lambda}| < 1$ para todo σ, ou seja, o método é **incondicionalmente estável.** Algumas vezes, esse fato é usado para justificar a utilização de um método implícito, pois sendo esse método incondicionalmente estável, não devemos nos preocupar com a amplificação de erros e portanto podemos utilizar uma malha menos fina para obter a solução.

Como nos casos anteriores, a análise da estabilidade pode ser feita também pelo critério da matriz, aplicado à equação de diferemças,

$$\mathbf{e}_j = A^{-1}\mathbf{e}_{j-1} + A^{-1}\tau_j \ ,$$

de maneira que teremos estabilidade se os autovalores de A^{-1} estiverem no disco unitário. Ver exercício (13.13).

Método de Crank-Nicolson

Este é um dos métodos mais utilizados na solução de equações parabólicas. Como no caso do método do Trapézio para equações diferenciais ordinárias, este método é a "média aritmética" do explícito com o implícito. Tomando pois a média entre as expressões (13.17) e (13.36) obtemos o **Método de Crank-Nicolson**:

$$U_{i,j+1} = U_{i,j} + \frac{\sigma(U_{i-1,j} + U_{i-1,j+1} - 2(U_{i,j} + U_{i,j+1}) + U_{i+1,j} + U_{i+1,j+1})}{2} \ , \qquad (13.39)$$

ou ainda,

$$\frac{U_{i,j+1} - U_{i,j}}{k} = \frac{a}{2h^2}\left(U_{i-1,j} - 2U_{i,j} + U_{i+1,j} + U_{i-1,j+1} - 2U_{i,j+1} + U_{i+1,j+1}\right) \ ,$$

cuja mol'ecula computacional é:

Figura 13.5: Molécula computacional do método de Crank-Nicolson.

Podemos observar que estas equações formam o seguinte sistema linear tridiagonal $A\mathbf{U}_{j+1} = B\mathbf{U}_j + \mathbf{c}_j$ que é diagonalmente dominante:

$$\begin{pmatrix} 2+2\sigma & -\sigma & 0 & \dots & 0 \\ -\sigma & 2+2\sigma & -\sigma & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & -\sigma & 2+2\sigma & -\sigma \\ 0 & \dots & 0 & -\sigma & 2+2\sigma \end{pmatrix} \begin{pmatrix} U_{1,j+1} \\ U_{2,j+1} \\ \vdots \\ U_{N-2,j+1} \\ U_{N-1,j+1} \end{pmatrix} =$$

$$\begin{pmatrix} 2-2\sigma) & \sigma & 0 & \dots & 0 \\ \sigma & 2-2\sigma & \sigma & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & \dots & \sigma & 2-2\sigma & \sigma \\ 0 & \dots & 0 & \sigma & 2-2\sigma \end{pmatrix} \begin{pmatrix} U_{1,j} \\ U_{2,j} \\ \vdots \\ U_{N-2,j} \\ U_{N-1,j} \end{pmatrix} + \begin{pmatrix} \sigma U_{0,j} \\ 0 \\ \vdots \\ 0 \\ \sigma U_{N,j} \end{pmatrix} ,$$

ou utilizando notação vetorial podemos escrever:

$$A\mathbf{U}_{j+1} = B\mathbf{U}_j + \mathbf{c}_j \tag{13.40}$$

Para se obter a solução em cada estágio, é preciso resolver um sistema tridiagonal. Note que, sendo A diagonalmente dominante, o problema discreto tem solução única.

Erro de Truncamento Local

Analogamente às definições de erro de truncamento local dos demais métodos definimos $\tau_{i,j}$ por:

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{\alpha}{2h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j} + u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}) + \tau_{i,j} \ .$$

Expandindo em série de Taylor em torno do ponto $(x_i, t_{j+\frac{1}{2}k})$, obtemos:

$$\frac{u_{i,j+1} - u_{i,j}}{k} = u_t + \left(\frac{k}{2}\right)^2 \frac{1}{6} u_{ttt} + O(k^4) \ ,$$

$$u_{i-1,j} - 2u_{i,j} + u_{i+1,j} = h^2 u_{xx} + \frac{h^4}{12} u_{xxxx} + O(h^6) \ ,$$

$$u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1} = h^2 u_{xx} + \frac{h^4}{12} u_{xxxx} \frac{h^2 k^2}{8} u_{xxtt} + \frac{h^4 k^2}{96} u_{xxxxtt} + O(h^6) \ .$$

Logo,

$$\tau_{i,j} = \frac{k^2}{24} u_{ttt} - \frac{\alpha h^2}{12} u_{xxxx} - \frac{\alpha k^2}{8} u_{xxtt} - a\frac{h^2 k^2}{96} u_{xxxxtt} + O(h^4) + O(k^4) = O(h^2 + k^2) \ .$$

Estabilidade

A análise da estabilidade segue o mesmo raciocínio desenvolvido para o método explícito. Substituindo (13.25) em (13.39) obtemos:

$$e^{\lambda} = \frac{1 - 2\sigma \operatorname{sen}^2 \frac{\beta}{2}}{1 + 2\sigma \operatorname{sen}^2 \frac{\beta}{2}} \ , \tag{13.41}$$

e concluímos que o método é **incondicionalmente estável**, pois $|e^{\lambda}|$ é sempre menor do que 1.

É possível provar que o erro de global satisfaz a equação vetorial $A\mathbf{e}_j = B\mathbf{e}_{j-1} + \tau_j$ e como A é inversível, tem-se estabilidade se todos os autovalores da matriz de amplificação $C = A^{-1}B$ estão no disco unitário. Ver exercício (13.15).

Condição de Fronteira de Neumann ou com Derivadas

Quando as condições de fronteira envolvem a derivada da função $u(x,t)$, dizemos tratar-se de **Condições de Fronteira de Neumann**. Podemos ter uma das condições de fronteira ou ambas, com derivadas. Teremos então um problema como a seguir:

$$\begin{aligned} u_t &= au_{xx} \ , \quad a > 0 \ , \quad 0 \leq x \leq L \ , \quad 0 < t < T \ , \\ u(x,0) &= \psi(x) \ , \quad 0 \leq x \leq L \ , \\ \frac{\partial u(0,t)}{\partial x} &= f(t) \ , \quad 0 < t < T \ , \\ \frac{\partial u(L,t)}{\partial x} &= g(t) \ , \quad 0 < t < T \ . \end{aligned} \tag{13.42}$$

Ao discretizarmos esse problema por qualquer das técnicas anteriores teremos agora que resolver a questão de que os valores $U_{0,j}$ e $U_{N,j}$ deixaram de ser conhecidos, veja a Figura 13.7.

Figura 13.7: Discretização do Problema de Neumann com inclusão de pontos fantasmas

Note que conhecemos os valores das derivadas direcionais sobre as fronteiras $x = 0$ e $x = L$ e não mais o valor da função como no caso da condição de fronteira de Dirichlet, de forma que devemos tratar esses valores $(U_{0,j}, U_{N,j})$ como incógnitas, ou seja, nossa discretização deve incluir os valores da função nos pontos $x_0 = 0$ e $x_N = L$ como incógnitas. O problema é que ao escrevermos uma equação para o ponto $x_0 = 0$, por exemplo, no caso do método explícito obtemos:

$$U_{0,j+1} = \sigma U_{-1,j} + (1 - 2\sigma)U_{0,j} + \sigma U_{1,j} \ ,$$

que contém o ponto $U_{-1,j}$ que não faz parte da malha. Devemos portanto utilizar a condição de fronteira para calculá-lo, fazendo:

$$f(t_j) = \frac{\partial u(0,t_j)}{\partial x} \simeq \frac{U_{1,j} - U_{-1,j}}{2h} \ , \tag{13.43}$$

onde foi utlizado a discretização da derivada por diferenças centrais. A equação (13.43) pode ser reescrita na forma:

$$U_{-1,j} = U_{1,j} - 2hf(t_j) \ , \tag{13.44}$$

e portanto o valor de $U_{-1,j}$ pode ser substituído por (13.44) na expressão do método numérico sem maiores problemas. Devemos no entanto observar que a notação vetorial introduzida em (13.28) deve ser modificada para refletir o fato de que $U_{0,j}$ e $U_{N,j}$ são agora incógnitas, assim:

$$\mathbf{U}_j = (U_{0,j}, U_{1,j}, \ldots, U_{N,j})^T \ ,$$

e portanto um vetor de $N+1$ componentes. Certamente a matriz A da equação (13.29) deve ser de ordem $(N+1)\times(N+1)$ e ter sua primeira e última linhas modificadas. Assim a equação equivalente a (13.29) para o caso de condições de fronteira com derivadas é:

$$\mathbf{U}_{j+1} = A\mathbf{U}_j + \mathbf{c}_j \ ,$$

onde A é a matriz **não simétrica**:

$$A = \begin{pmatrix} 1-2\sigma & 2\sigma & 0 & \ldots & 0 \\ \sigma & 1-2\sigma & \sigma & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & \sigma & 1-2\sigma & \sigma \\ 0 & \ldots & 0 & 2\sigma & 1-2\sigma \end{pmatrix} \ , \tag{13.45}$$

e $\mathbf{c}_j$ é um vetor de $(N+1)$ componentes contendo informações das condições de fronteira dado por:

$$\mathbf{c}_j = (-2hf(jk), 0, \ldots, 0, -2hg(jk))^T \ .$$

Já para o caso de um método implícito teremos um pouco mais de dificuldades. Consideremos o método de Crank-Nicolson (13.39) que pode ser reescrito na forma:

$$-\sigma U_{i+1,j+1} + (2+2\sigma)U_{i,j+1} - \sigma U_{i-1,j+1} = \sigma U_{i+1,j} + (2-2\sigma)U_{i,j} + \sigma U_{i-1,j} \ . \tag{13.46}$$

Novamente quando $i = 0$ ou $i = N$ teremos o aparecimento dos termos $U_{-1,j+1}$, $U_{-1,j}$, $U_{N+1,j+1}$ e $U_{N+1,j}$ que não fazem parte da malha e portanto são pontos fantasmas. Da mesma maneira que fizemos para o caso explícito em (13.44) eliminamos esses valores da equação (13.46), para obter a equação matricial:

$$A\mathbf{U}_{j+1} = B\mathbf{U}_j + \mathbf{c}_j \ ,$$

onde as matrizes A e B são de ordem $N+1$ e dadas por:

$$A = \begin{pmatrix} 2+2\sigma & -2\sigma & 0 & \ldots & 0 \\ -\sigma & 2+2\sigma & -\sigma & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & -\sigma & 2+2\sigma & -\sigma \\ 0 & \ldots & 0 & -2\sigma & 2+2\sigma \end{pmatrix} \ ,$$

$$B = \begin{pmatrix} 2-2\sigma) & 2\sigma & 0 & \ldots & 0 \\ \sigma & 2-2\sigma & \sigma & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & \sigma & 2-2\sigma & \sigma \\ 0 & \ldots & 0 & 2\sigma & 2-2\sigma \end{pmatrix} \ ,$$

e o vetor independente $\mathbf{c}_j = (2h\sigma(f(t_{j+1}) - f(t_j)), 0, \ldots, 0, 2h\sigma(g(t_{j+1}) - g(t_j)))^T$.

Observações:

a) A matriz A do método de Crank-Nicolson para problemas com condição de fronteira de Neumann não é simétrica, mas continua estritamente diagonalmente dominante. A perda da simetria é um fator de considerável complicação para a solução do sistema linear.

b) No caso do domínio possuir uma fronteira irregular, os pontos próximos da fronteira precisam de tratamento específico, por meio de interpolação. Este aspecto será tratado com mais detalhe nas equações elípticas.

c) É também possível encontrar na prática problemas com condições de fronteira do tipo misto, ou seja, $u_x(0,t) - \alpha_1 u(0,t) = g(t)$. O tratamento numérico desse tipo de condição de fronteira é uma combinação daquele dos casos de fronteira de Dirichlet e de Neumann estudados acima.

13.4 Problemas Não Lineares

A maioria dos métodos discutidos anteriormente podem ser generalizados para equações lineares com coeficientes variáveis, ou seja, $\alpha = \alpha(x,t)$ na equação (13.14). Mas encontramos algumas dificuldades como:

- embora as matrizes continuem diagonalmente dominantes, σ não é mais constante;
- temos que calcular $\alpha_{i,j}$ a cada passo, e no método de Crank-Nicolson, por exemplo, precisamos de $\alpha_{i,j}$ no ponto intermediário, para isso calculamos:

$$\alpha(x_i, t_{j+\frac{k}{2}}) \quad \text{ou} \quad \frac{\alpha(x_i,t_j) + \alpha(x_i,t_{j+1})}{2} \ ;$$

- não poderemos usar análise de Fourier para fazer a análise da estabilidade de maneira global, apenas localmente.

Quando $\alpha = \alpha(x,t,u)$ ou $\alpha = \alpha(u)$ temos o chamado problema quase-linear, neste caso o método explícito pode ser usado sem dificuldades, pois sua utilização não requer a solução de equações não lineares. Já para os métodos implícito e de Crank-Nicolson obtemos um sistema de equações não lineares, que podemos resolver por aproximações sucessivas ou pelo método de Newton, com a aproximação inicial calculada pelo método explícito. Na prática, no entanto, recomenda-se a utilização de métodos implícitos para evitar problemas de estabilidade, uma vez que estes são incondicionalmente estáveis.

O caso mais geral de uma equação não linear parabólica é representado pela expressão:

$$f(x,t,u,u_t,u_x,u_{xx}) = 0 \ . \tag{13.47}$$

Casos particulares importantes são representados pela **equação de Burgers**, onde se tem $f(x,t,u,u_t,u_x,u_{xx}) = u_t + uu_x - \alpha u_{xx} = 0$ e pela **equação de Ritchmyer** ([?]) onde $f(x,t,u,u_t,u_x,u_{xx}) = u_t - (u^m)_{xx} = 0, \quad m \geq 2$.

Um método explícito para resolver (13.47), simplesmente avalia essa expressão no ponto (x_i, t_j) e substitui a derivada em t por diferença progressiva e as derivadas em x por diferenças centrais. No caso em que a derivada temporal pode ser escrita de forma explícita em função das outras variáveis, ou seja quando a equação é da forma:

$$u_t = f(x,t,u,u_x,u_{xx}) \ ,$$

podemos discretizá-la facilmente pelo método explícito para obter:

$$U_{i,j+1} = U_{i,j} + kf(x_i, t_j, U_{i,j}, \frac{U_{i+1,j} - U_{i-1,j}}{2h}, \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}) \ . \tag{13.48}$$

Uma análise da estabilidade linear desse método é possível e pode ser encontrada em detalhes em [**?**] páginas 74-76. O resultado demonstrado em [**?**] é o seguinte:

Teorema 13.4 *Seja f uma função satisfazendo as seguintes hipóteses:*

1. $f_{u_{xx}} \geq \gamma > 0$,
2. $|f_u| + |f_{u_x}| + f_{u_{xx}} \leq \beta$,

onde f_ϕ representa a derivada parcial da função f com relação ao argumento ϕ e supomos que u seja uma função 4 vezes diferenciável em x e 2 vezes em t. Então o método (13.48) é convergente se:

$$h \leq \frac{2\gamma}{\beta} \ \text{ e } \ 0 < \frac{k}{h^2} \leq \frac{1 - \beta k}{2\beta} \ .$$

No caso em que:

$$f(x, t, u, u_x, u_{xx}) = (u^m)_{xx} = m(m-1)u^{m-2}\left(\frac{\partial u}{\partial x}\right)^2 + mu^{m-1}\frac{\partial^2 u}{\partial x^2}, \quad m \geq 2, \tag{13.49}$$

teremos $0 < \gamma = mu^{m-1}$ e $\beta = mu^{m-1} + 2m(m-1)|u^{m-2}u_x| + m(m-1)(m-2)|u^{m-3}(u_x)^2|$.

Observe que assumimos que $\gamma > 0$ e portanto $\beta > \gamma$ de forma que o teorema acima impõe as condições de estabilidade:

1. $h < 2\gamma/\beta < 2$
2. $\dfrac{k}{h^2} < (1-\beta k)/(2\beta)$ ou seja $\beta\dfrac{k}{h^2} < (1-\beta k)/2 < 1/2$. Usando agora que $\beta > \gamma$ na última desigualdade obtemos:
$$mu^{m-1}\frac{k}{h^2} = \gamma\frac{k}{h^2} < \frac{1}{2} \ .$$

Esse mesmo resultado poderia ser obtido por comparação com a equação linear com coefientes constantes reescrevendo a equação (13.49) na forma

$$u_t = (mu^{m-1}u_x)_x \ ,$$

e compararando com a equação com coeficientes constantes, $u_t = (\alpha u_x)_x$, para concluir que a condição de estabilidade do método explícito toma a forma:

$$\sigma = \frac{\alpha k}{h^2} = \frac{mu^{m-1}k}{h^2} \leq \frac{1}{2} \ .$$

Isto quer dizer que para problemas não lineares a estabilidade depende, além da equação de diferenças finitas, da solução do problema que está sendo resolvido e portanto a condição de estabilidade pode variar ao longo do domínio. Para obter-se um método implícito substitui-se as derivadas da variável espacial por fórmulas envolvendo o mesmo nível de tempo. Um exemplo de método implícito para resolver (13.47) é o método de Crank-Nicolson dado por:

$$f\left(x_i, t_{j+\frac{1}{2}}, \frac{U_{i,j+1} + U_{i,j}}{2}, \frac{U_{i,j+1} - U_{i,j}}{k}, \frac{\mu\delta_x U_{i,j+1} + \mu\delta_x U_{i,j}}{2h}, \frac{\delta_x^2 U_{i,j+1} + \delta_x^2 U_{i,j}}{2h^2}\right) = 0 \ ,$$

onde os operadores μ e δ_x estão definidos em (**??**). A equação acima deve ser aplicada em todos os pontos do domínio para os quais quer-se calcular uma aproximação, produzindo um sistema de equações não lineares nas variáveis $U_{i,j}$. Esse sistema pode ser resolvido por aproximações sucessivas ou pelo método de Newton, (veja Capítulo 4).

13.5 Equações Parabólicas em Duas Dimensões

Nesta seção consideraremos a discretização de equações parabólicas lineares em duas dimensões, da forma:

$$u_t = \alpha_1 u_{xx} + \alpha_2 u_{yy} \ ,$$

onde u, α_1 e α_2 são funções de x, y e t; x e y variáveis espaciais e t variável tempo, ou em notação de operadores,

$$u_t \ = \ Lu \ , \tag{13.50}$$

com $L \equiv \alpha_1 u_{xx} + \alpha_2 u_{yy}$.

Todos os métodos discutidos anteriormente podem ser adaptados para este caso mas o esforço computacional aumenta quadraticamente, de forma que a simples adaptação não é uma solução satisfatória.

Apresentamos a seguir alguns métodos criados para contornar esse problema, exigindo um número menor de operações aritméticas.

Inicialmente supomos que a região de definição do problema seja formada pelo retângulo $[0, a] \times [0, b]$ do plano $x - y$ e pelo intervalo de tempo $[0, \infty)$. Essa regi ao do plano é coberta por uma malha retangular com lados paralelos aos eixos, com espaçamento h_1 na direção x, h_2 na direção y e por uma discretização temporal com passo k. Os pontos da malha (x, y, t) serão dados por $x = lh_1, y = mh_2$ e $t = nk$, onde l, m, n são inteiros. Uma função discreta definida nessa malha será denotada por $U^n_{l,m}$.

Da expansão em série de Taylor de $u(x, y, t_{n+1}) = u(x, y, t_n + k)$ em torno do ponto (x, y, t_n), obtemos:

$$u(x, y, t_{n+1}) = u(x, y, t_n) + ku_t(x, y, t_n) + \frac{k^2}{2} u_{tt}(x, y, t_n) + \cdots \ . \tag{13.51}$$

Observando agora que $u_t = Lu$ e que L independe de t, deduzimos por diferenciação com relação a t que $u_{tt} = (Lu)_t = Lu_t = L \times Lu = L^2 u$. Assim podemos mostrar por indução que:

$$\frac{\partial^p u}{\partial t^p} = L^p u \ .$$

Levando esse resultado em (13.51), obtemos:

$$\begin{aligned} u(x, y, t_{n+1}) \ &= \ u(x, y, t_n) + kLu(x, y, t_n) + \frac{k^2}{2} L^2 u(x, y, t_n) + \cdots \\ &= \ (I + kL + \frac{k^2}{2} L^2 + \frac{k^3}{3!} L^3 + \cdots) u(x, y, t_n) = exp(kL) u(x, y, t_n) \ . \end{aligned} \tag{13.52}$$

Avaliando a expressão (13.52) no ponto (x_l, y_m, t_n) obtemos a fórmula (13.53), que será utilizada na dedução dos diversos métodos a seguir.

$$u^{n+1}_{l,m} = exp(kL) u^n_{l,m} \ . \tag{13.53}$$

Método Explícito

Vamos exemplificar esta técnica usando

$$L \equiv u_{xx} + u_{yy} = D_1^2 + D_2^2, \tag{13.54}$$

ou seja, $\alpha_1 \equiv \alpha_2 \equiv 1$ e $D_1 = u_x$, $D_2 = u_y$.

A equação (13.53) torna-se:

$$u^{n+1} = \exp(kD_1^2)\exp(kD_2^2)u^n$$

onde $u^n = u^n_{l,m}$.

Como

$$D_1^2 = \frac{1}{h^2}(\delta_x^2 - \frac{1}{12}\delta_x^4 + \frac{1}{90}\delta_x^6 \cdots)$$

e

$$D_2^2 = \frac{1}{h^2}(\delta_y^2 - \frac{1}{12}\delta_y^4 + \frac{1}{90}\delta_y^6 \cdots),$$

veja exercício (**??**). Então

$$u^{n+1} = [1 + \sigma\delta_x^2 + \frac{1}{2}\sigma(\sigma - \frac{1}{6})\delta_x^4 \cdots][1 + \sigma\delta_y^2 + \frac{1}{2}\sigma(\sigma - \frac{1}{6})\delta_y^4 \cdots]u^n, \tag{13.55}$$

onde $\sigma = k/h^2$. Aqui estamos utilizando $h_1 = h_2 = h$, como espaçamento nas variáveis x e y. A conclusão final não será muito distinta se considarmos espaçamentos diferentes para cada uma das variáveis espaciais. No entanto, alertamos que essa talvez seja uma situação mais realista para aplicações práticas.

Vários métodos explícitos podem ser obtidos da equação (13.55). Por exemplo, multiplicando as duas séries e em seguida considerando somente os termos de primeira ordem obtemos,

$$U^{n+1}_{l,m} = [1 + \sigma(\delta_x^2 + \delta_y^2)]U^n_{l,m} + O(k^2 + kh^2), \tag{13.56}$$

que é o método explícito padrão envolvendo cinco pontos no nível de tempo $t = nk$.

Outro método simples pode ser obtido da equação (13.55) considerando os termos de primeira ordem em cada uma das expansões separadamente

$$U^{n+1}_{l,m} = (1 + \sigma\delta_x^2)(1 + \sigma\delta_y^2)U^n_{l,m} + O(k^2 + kh^2), \tag{13.57}$$

Esta fórmula envolve nove pontos do nível de tempo $t = nk$.

Estabilidade

Analogamente ao caso unidimensional temos o critério de von Neumann e o critério da matriz para análise da estabilidade. No entanto, em duas dimensões a análise da estabilidade pelo método da matriz se torna um tanto complexa, pois na tentativa de generalizar o procedimento feito para o problema unidimensional por exemplo para a equação (**??**) obtemos um sistema linear $\mathbf{U}^{n+1} = A\mathbf{U}^n + \mathbf{c}^n$, onde agora, $\mathbf{U}^n$ é um vetor de vetores e A uma matriz de matrizes. Se compararmos com a matriz (13.30) veremos que no lugar dos elementos dessa matriz aparecerá agora matrizes. Este fato, complica a análise do problema de autovalores.

Pela sua simplicidade e facilidade de exposição concentraremos aqui no estudo do critério de von Neumann que assume uma decomposição harmônica dos erros em um dado nível de tempo, por exemplo, $t = 0$. Um modo individual, representando a decomposição é então dado por:

$$e^{\alpha t}e^{I\beta x}e^{I\gamma y} \tag{13.58}$$

onde β e γ são números reais arbitrários e $\alpha \equiv \alpha(\beta, \gamma)$ é em geral complexo. Tomando essa solução nos pontos da malha obtemos:

$$e^{\alpha t_n} e^{I\beta x_i} e^{I\gamma y_j} = (e^{\alpha k})^n e^{I\beta x_i} e^{I\gamma y_j}$$

E portanto, essa componente do erro será uniformente limitada se:

$$|e^{\alpha k}| \leq 1$$

para todo α.

Se (13.58) é substituído na fórmula (13.56), o resultado eliminando-se fatores comuns torna-se

$$e^{\alpha k} = 1 - 4\sigma \left(\text{sen}^2 \frac{\beta h}{2} + \text{sen}^2 \frac{\gamma h}{2} \right).$$

Para estabilidade $|e^{\alpha k}| \leq 1$ e, então

$$-1 \leq 1 - 4\sigma \left(\text{sen}^2 \frac{\beta h}{2} + \text{sen}^2 \frac{\gamma h}{2} \right) \leq 1.$$

O lado direito da desigualdade é satisfeito, e o lado esquerdo resulta em:

$$\sigma \leq \frac{1}{2\left(\text{sen}^2 \frac{\beta h}{2} + \text{sen}^2 \frac{\gamma h}{2} \right)} \leq \frac{1}{4},$$

então temos a condição de estabilidade $\sigma \leq \frac{1}{4}$.

Muitas vezes a restrição imposta pela estabilidade sobre o passo temporal, torna-se muito restritiva, e pode levar o usuário a preferir a utilização de um método implícito para escapar dessa restrição. Na próxima seção apresentamos brevemente os métodos mais utilizados na prática.

Métodos de Direções Alternadas Implícitos

A discretização de uma equação parabólica em duas dimensões por um método implícito leva à necessidade de solução de um conjunto de sistemas de equações lineares cuja matriz tem a dimensão do número de pontos da malha nas variáveis espaciais, isto é se tivermos N pontos na direção x e M na direção y, a cada passo de tempo, teremos que resolver um sistema linear com NM equações. Esse processo pode ser extremamente caro do ponto de vista computacional se não levarmos em consideração a estrutura muito especial da matriz dos coeficientes. Por exemplo a discretização implícita equivalente a (13.56) é:

$$U_{i,j}^{n+1} = U_{i,j}^{n} + \sigma(\delta_x^2 + \delta_y^2) U_{i,j}^{n+1} \tag{13.59}$$

ou seja,

$$(1 + \sigma(\delta_x^2 + \delta_y^2)) U_{i,j}^{n+1} = U_{i,j}^{n}$$

que resulta num sistema linear cuja matriz A tem no máximo 5 elementos não nulos por linha. Na maioria das vezes é possível arranjar as incógnitas desse sistema de tal forma que os 5 elementos não nulos de cada linha estejam posicionados de maneira a formar uma matriz com 5 diagonais, a principal, uma imediatamente acima e outra imediatamente abaixo desta, e duas outras localizadas a certa distância da diagonal principal. Obviamente esta é uma matriz muito especial e não devemos tentar resolver o sistema linear resultante sem ter essa característica em mente. A grande dificuldade é que não existem métodos especiais eficientes para tratar o problema com cinco diagonais, e o método de Gauss não é muito adequado pois ao aplicarmos o processo de triangularização elementos que eram nulos originalmente o deixam de ser ao longo do processo, provocando o processo conhecido como "fill in". Já esta mesma dificuldade não ocorre se a matriz tiver apenas 3 diagonais; a principal e as duas adjacentes. Nesse caso

o método de eliminação de Gauss pode ser aplicado sem nenhuma dificuldade. A tentativa de resolver o problema bidimensional resolvendo-se apenas sistemas tridiagonais é materializada pela concepção dos Métodos de Direções Alternadas (ADI). Métodos ADI são métodos de 2-passos onde em cada passo apenas uma das variáveis é tratada implicitamente. No primeiro passo u_{xx} é discretizado implicitamente e u_{yy} é tratado explicitamente, no segundo passo os papeis se invertem e assim sucessivamente. O esforço computacional do método ADI é cerca de tres vezes o do método explícito, preço que pagamos ao utilizar um método incondicionalmente est'avel.

Ilustraremos estes métodos com respeito a equação (13.50) com L como em (13.54). A região a ser considerada consiste em R x $[t \geq 0]$, onde R é uma região arbitrária fechada em $\mathbb{R}^2$. Inicialmente tomamos R como quadrado $[0 \leq x \leq 1, 0 \leq y \leq 1]$.

Da equação (13.53) temos:

$$u^{n+1} = \exp(k(D_1^2 + D_2^2))u^n$$

$$u^{n+1} = \exp(\frac{k}{2}(D_1^2 + D_2^2) + \frac{k}{2}(D_1^2 + D_2^2))u^n$$

logo,

$$\exp(-\frac{k}{2}(D_1^2 + D_2^2))u^{n+1} = \exp(\frac{k}{2}(D_1^2 + D_2^2))u^n \qquad (13.60)$$

$$\exp(-\frac{k}{2}D_1^2)\exp(-\frac{k}{2}D_2^2)u^{n+1} = \exp(\frac{k}{2}D_1^2)\exp(\frac{k}{2}D_2^2)u^n,$$

cuja expansão e truncamento da série fornece a equação:

$$(1 - \frac{1}{2}\sigma\delta_x^2)(1 - \frac{1}{2}\sigma\delta_y^2)U^{n+1} = (1 + \frac{1}{2}\sigma\delta_x^2)(1 + \frac{1}{2}\sigma\delta_y^2)U^n + O(k^3 + kh^2), \qquad (13.61)$$

este método é uma modificação do método de Crank-Nicolson em duas dimensões que é dado por:

$$(1 - \frac{1}{2}\sigma\delta_x^2 - \frac{1}{2}\sigma\delta_y^2)U^{n+1} = (1 + \frac{1}{2}\sigma\delta_x^2 + \frac{1}{2}\sigma\delta_y^2)U^n + O(k^3 + kh^2). \qquad (13.62)$$

Note que para obter (13.61) de (13.62) um termo da forma $\frac{\sigma}{4}\delta_x^2\delta_y^2$ foi adicionado em ambos os membros de (13.62).

A equação (13.61) pode ser interpretada de uma forma mais conveniente para a implementação computacional introduzindo-se um passo intermediário para "decompor" (13.61) em duas equações:

$$\begin{cases} (1 - \frac{1}{2}\sigma\delta_x^2)U^{n+1*} & = & (1 + \frac{1}{2}\sigma\delta_y^2)U^n \\ (1 - \frac{1}{2}\sigma\delta_y^2)U^{n+1} & = & (1 + \frac{1}{2}\sigma\delta_x^2)U^{n+1*} \end{cases} \qquad (13.63)$$

ou seja, o passo intermediário representa uma solução na direção x e o passo final uma solução na direção y.

O método decomposto (13.63) com $U^{n+1^*} = U^{n+1/2}$, isto é o passo intermediário é interpretado como um "meio" passo, foi introduzido por Peaceman e Rachford [**?**] e é conhecido como *método de Peaceman e Rachford.*

Um método decomposto com precisão mais alta pode ser obtido da equação (13.60) substituindo D_1^2 e D_2^2 por, (veja exercício (**??**)).

$$D_1^2 = \frac{\delta_x^2}{h^2(1 + \frac{1}{12}\delta_x^2)} \quad \text{e} \quad D_2^2 = \frac{\delta_y^2}{h^2(1 + \frac{1}{12}\delta_y^2)}$$

e expandindo para obter:

$$[1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_x^2][1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_y^2]U^{n+1}=[1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_x^2][1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_y^2]U^n+O(k^3+kh^4), \quad (13.64)$$

que pode ser decomposto em duas equações:

$$\begin{cases} (1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_x^2)U^{n+1^*} &= (1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_y^2)U^n \\ (1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_y^2)U^{n+1} &= (1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_x^2)U^{n+1^*} \end{cases} \quad (13.65)$$

este método foi obtido por Mitchell e Fairweather [?].

As equações (13.63) e (13.65) são exemplos de métodos envolvendo a solução de sistemas tridiagonal ao longo das linhas paralelas aos eixos x e y respectivamente. Estes são os métodos ADI.

As fórmulas (13.61) e (13.64) podem ser decompostas de uma outra maneira sugerida por D'Yakonov [?]:

$$\begin{cases} (1-\frac{1}{2}\sigma\delta_x^2)U^{n+1^*} &= (1+\frac{1}{2}\sigma\delta_x^2)(1+\frac{1}{2}\sigma\delta_y^2)U^n \\ (1-\frac{1}{2}\sigma\delta_y^2)U^{n+1} &= U^{n+1^*}, \end{cases}$$

e

$$\begin{cases} (1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_x^2)U^{n+1^*} &= (1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_x^2)(1+\frac{1}{2}(\sigma+\frac{1}{6})\delta_y^2)U^n \\ (1-\frac{1}{2}(\sigma-\frac{1}{6})\delta_y^2)U^n &= U^{n+1^*} \end{cases}$$

respectivamente.

Finalmente, Douglas e Rachford [?] formularam um método ADI que é dado na forma decomposta por:

$$\begin{cases} (1-\sigma\delta_x^2)U^{n+1^*} &= (1+\sigma\delta_y^2)U^n \\ (1-\sigma\delta_y^2)U^{n+1} &= U^{n+1^*}-\sigma\delta_y^2U^n, \end{cases}$$

e é conhecido como o *método Douglas-Rachford.* Eliminando-se U^{n+1^*} temos a fórmula:

$$(1-\sigma\delta_x^2)(1-\sigma\delta_y^2)U^{n+1}=(1+\sigma^2\delta_x^2\delta_y^2)U^n,$$

que pode ser decomposta, de acordo com o método de D'Yakonov, em:

$$\begin{cases} (1-\sigma\delta_x^2)U^{n+1^*} &= (1+\sigma^2\delta_x^2\delta_y^2)U^n, \\ (1-\sigma\delta_y^2)U^{n+1} &= U^{n+1^*}. \end{cases}$$

Usando o método de von Neumann, mostra-se a estabilidade, dos métodos ADI apresentados nesta seção, para todo valor de $\sigma>0$, ver exercício (13.25).

Método Localmente Unidimensional

Vamos ilustrar os métodos localmente um-dimensionais (LOD) resolvendo a equação

$$u_t=u_{xx}+u_{yy},$$

que pode ser reescrita como o par de equações

$$\frac{1}{2}u_t=u_{xx} \quad \text{e} \quad \frac{1}{2}u_t=u_{yy}. \quad (13.66)$$

A idéia é portanto, aproximar a solução de um problema em duas dimensões resolvendo dois problemas unidimensionais que localmente representam o problema original. As discretizações explícitas mais simples dessas fórmulas são:

$$\begin{aligned}\frac{1}{2}\left(\frac{U^{n+\frac{1}{2}}-U^n}{k/2}\right) &= \frac{\delta_x^2}{h^2}U^n\\ \frac{1}{2}\left(\frac{U^{n+1}-U^{n+\frac{1}{2}}}{k/2}\right) &= \frac{\delta_y^2}{h^2}U^{n+\frac{1}{2}}\end{aligned}$$

que pode ser reescrita na forma compacta como:

$$U^{n+\frac{1}{2}}=(1+\sigma\delta_y^2)U^n \quad \text{e} \quad U^{n+1}=(1+\sigma\delta_x^2)U^{n+\frac{1}{2}}, \tag{13.67}$$

e eliminando $U^{n+\frac{1}{2}}$ temos:

$$U^{n+1}=(1+\sigma\delta_x^2)(1+\sigma\delta_y^2)U^n.$$

Obtemos assim a equação (13.57) que aproxima a solução com erro de truncamento da $O(k^2+kh^2)$. Para um problema de valor inicial com condições dadas sobre o plano $t=0$, $-\infty<x<\infty, -\infty<y<\infty$, se usarmos o método LOD (13.67) temos a mesma precisão e estabilidade mas mais economia de cálculos do que usando (13.57). Já o método de Crank-Nicolson para o par de equações (13.66) toma a forma:

$$\begin{aligned}\frac{1}{2}\left(\frac{U^{n+\frac{1}{2}}-U^n}{k/2}\right) &= \frac{\delta_x^2U^{n+\frac{1}{2}}+\delta_x^2U^n}{2h^2}\\ \frac{1}{2}\left(\frac{U^{n+1}-U^{n+\frac{1}{2}}}{k/2}\right) &= \frac{\delta_y^2U^{n+1}+\delta_y^2U^{n+\frac{1}{2}}}{2h^2}\end{aligned}$$

que pode ser reescrito como:

$$\begin{aligned}\left(1-\frac{\sigma}{2}\delta_x^2\right)U^{n+\frac{1}{2}} &= \left(1+\frac{\sigma}{2}\delta_x^2\right)U^n\\ \left(1-\frac{\sigma}{2}\delta_y^2\right)U^{n+1} &= \left(1+\frac{\sigma}{2}\delta_y^2\right)U^{n+\frac{1}{2}}\end{aligned} \tag{13.68}$$

Se os operadores δ_x^2 e δ_y^2 comutam, e este é o caso quando o domínio é um retângulo de lados paralelos aos eixos x e y, então o método (13.68) é equivalente ao método Peaceman e Rachford. O método LOD construído acima é de segunda ordem, é incondicionalmente estável e envolve apenas a solução de sistemas tridiagonais.

Neste capítulo tentamos apresentar uma coleção representativa dos diferentes tipos de métodos e problemas em equações parabólicas. No entanto o leitor com aplicações mais específicas pode encontrar um grande número delas em Ames [?], Thomas [?] Lapidus & Pinder [?] e Sod [?].

13.6 Equações Elípticas

Problemas de equilíbrio em duas ou três dimensões geralmente dão origem à equações elípticas. Exemplos típicos dessa classe são problemas de difusão e de pressão, problemas em elasticidade, problemas de camada limite, problemas de vibração de membranas, etc. Mais simplificadamente, os problemas elípticos

caracterizam-se pela propagação de suas propriedades físicas em todas as direções coordenadas indistintamente, ao contrário das equações parabólicas e hiperbólicas onde essas propriedades propagam-se em direções preferenciais. Daí porque as condições de fronteira de um problema elíptico são normalmente especificadas ao longo de toda a fronteira.

Seja R uma região limitada do plano com fronteira ∂R. A equação

$$a(x,y)\frac{\partial^2 u}{\partial x^2} + 2b(x,y)\frac{\partial^2 u}{\partial x \partial y} + c(x,y)\frac{\partial^2 u}{\partial y^2} = d\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right) \tag{13.69}$$

de acordo com a definição apresentada na seção 2.2 do capítulo 2 é elíptica em R se $b^2 - ac < 0$ para todo ponto (x, y) de R. Como já observado anteriormente uma equação diferencial como (13.69) necessita de condições iniciais e/ou de fronteira para constituir-se num *problema bem posto*. Isto ocorre porque, suponha que w seja uma solução da equação (13.69) com $d \equiv 0$, então $v = u + \alpha w$ será uma solução de (13.69) para qualquer valor de α, onde u é uma solução particular de (13.69). Esta observação indica a possibilidade de que a equação (13.69) sozinha tenha infinitas soluções e portanto precisamos de condições adicionais para assegurar unicidade.

Imaginamos que não seja dificil para o leitor compreender a necessidade de um problema ter uma única solução para que seu tratamento numérico possa ser considerado. Não faria muito sentido tentar aproximar a solução de um problema para o qual não existe uma, ou que tenha uma infinidade delas, caso em que o método numérico ficaria totalmente indeciso sobre qual delas perseguir. De forma que parece claro que só tem sentido tentar resolver numéricamente um problema com solução única. Já a importância da continuidade em relação aos dados iniciais, para a solução numérica, pode não ser tão óbvia. Ocorre que quando resolvemos um problema numéricamente, estamos fazendo aproximações, e tudo se passa como se na verdade estivessemos resolvendo um outro problema cujos dados iniciais sofreram uma pequena perturbação. Se o problema não se comporta bem com relação à pequenas perturbações nesses dados, a solução numérica obtida será desastrosa.

Vimos então a necessidade de que à equação (13.69) seja adicionada condições de fronteira e que essas condições de fronteira devem ser escolhidas de maneira a resultar em um problema bem posto. Felizmente, a maioria dos problemas que reclamam um tratamento numérico provêm de aplicações práticas e essas mesmas aplicações determinam as condições de fronteira, formando problemas bem postos.

Três tipos de problemas distintos envolvendo a equação (13.69) podem ser destacados dependendo das condições de fronteira:

- (i) o problema de Dirichlet, requer que a solução u de (13.69) seja conhecida sobre a fronteira ∂R, isto é,
$$u(x,y) = f(x,y), \quad (x,y) \in \partial R.$$
- (ii) quando $\frac{\partial u}{\partial n}$ é conhecida sobre ∂R, ou seja,
$$\frac{\partial u}{\partial n} = g(x,y), \quad (x,y) \in \partial R,$$
onde n é a normal externa à fronteira ∂R, o problema de valor de fronteira é conhecido como problema de Neumann.
- (iii) o problema de Robbins ou misto, surge quando conhecemos
$$\alpha(x,y)u + \beta(x,y)\frac{\partial u}{\partial n} = \gamma(x,y) \quad \text{sobre } \partial R,$$
onde $\alpha(x,y) > 0, \beta(x,y) > 0, \quad (x,y) \in \partial R.$

Quando em (13.69) tomamos $a = c \equiv 1$ e $b = d \equiv 0$ obtemos o protótipo de equação elíptica mais conhecido que é a famosa equação de Laplace

$$-\left(u_{xx} + u_{yy}\right) = 0. \tag{13.70}$$

Passamos à seguir a discutir métodos de aproximação para a classe de equações elípticas.

13.7 Métodos de Diferenças Finitas

Os métodos de diferenças finitas, a exemplo do que fizemos no capítulo 3 para as equações parabólicas, consistem em substituir as derivadas parciais presentes na equação diferencial por aproximações por diferenças finitas. Para isto é necessário que os pontos onde essas diferenças serão calculadas sejam pré estabelecidos, ou seja é necessário a definição de uma malha de pontos no domínio. Para ilustrar como esta discretização é realizada na prática consideremos a equação de Poisson:

$$-\Delta u = -(u_{xx} + u_{yy}) = f, \tag{13.71}$$

definida no retângulo $R = \{(x, y), 0 \leq x \leq a, 0 \leq y \leq b\}$ com condição de Dirichlet:

$$u = g(x, y) \tag{13.72}$$

sobre a fronteira, ∂R desse retângulo.

Primeiramente, para que possamos aproximar u_{xx} e u_{yy} por diferenças finitas cobrimos a região R com uma malha. Escolhemos a opção mais óbvia que consiste em traçar linhas paralelas aos eixos coordenados, conforme ilustrado na figura 4.1. Os pontos dessa malha serão denotados por $(x_i, y_j), x_i = ih, y_j = jk, i = 0, 1, \cdots M, j = 0, 1, \cdots N$, onde h representa o espaçamento na direção x e k na direção y. Denotamos por R_δ os pontos da malha interiores a R, isto é,

$$R_\delta = \{(x_i, y_j), 0 < i < M, 0 < j < N\}$$

e por ∂R_δ os pontos da malha que estão sobre a fronteira ou seja,

$$\partial R_\delta = \{(x_i, y_j), (i = 0, M, 0 \leq j \leq N) \text{ e } (0 \leq i \leq M, j = 0, N)\}$$

Figura 4.1: Malha de discretização

Podemos agora aproximar as derivadas da equação (13.71) da seguinte forma: A equação (13.71) é válida para todos os pontos de R então em particular para um ponto genérico de R_δ podemos escrever,

$$-(u_{xx}(x_i, y_j) + u_{yy}(x_i, y_j)) = f(x_i, y_j). \tag{13.73}$$

Assim, as derivadas podem ser aproximadas por:

$$u_{xx}(x_i, y_j) \simeq \frac{u(x_i + h, y_j) - 2u(x_i, y_j) + u(x_i - h, y_j)}{h^2}$$

$$u_{yy}(x_i, y_j) \simeq \frac{u(x_i, y_j + k) - 2u(x_i, y_j) + u(x_i, y_j - k)}{k^2}.$$

Substituindo essas aproximações em (13.73) obtemos:

$$-\left(\frac{u(x_i + h, y_j) - 2u(x_i, y_j) + u(x_i - h, y_j)}{h^2} + \frac{u(x_i, y_j + k) - 2u(x_i, y_j) + u(x_i, y_j - k)}{k^2}\right) \simeq f(x_i, y_j). \tag{13.74}$$

Note que a expressão (13.74) não representa uma equação porque o segundo membro é somente uma aproximação para o primeiro e portanto não temos uma igualdade. Isto decorreu de termos substituído $u_{xx}(x_i, y_j)$ e $u_{yy}(x_i, y_j)$ por suas respectivas aproximações. Podemos transformar (13.74) numa equação, simplesmente trocando o sinal $\simeq$ pelo de igualdade. Se assim procedermos, no entanto, não poderemos mais garantir que os valores numéricos presentes no lado esquerdo de (13.74) coincidam com os valores da solução de (13.71) nos mesmos pontos.

Seguindo a notação utilizada na literatura denotaremos por $u_{i,j}$ o valor da solução no ponto (x_i, y_j) e por $U_{i,j}$ a solução da equação de diferenças:

$$-\left(\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{k^2}\right) = f_{i,j}. \tag{13.75}$$

A equação (13.75) deverá ser aplicada para todos os pontos em R_δ. Para os pontos em ∂R_δ calculamos $U_{i,j}$ da condição de fronteira de Dirichlet

$$U_{i,j} = g(x_i, y_j). \tag{13.76}$$

Nossa esperança quando escrevemos a equação (13.75) é que $U_{i,j}$ seja uma aproximação para $u(x_i, y_j)$, isto é, $U_{i,j} \simeq u(x_i, y_j)$. Demonstraremos mais adiante que, de fato, isto é verdadeiro. Provaremos mais ainda que $U_{i,j}$ "converge" para $u(x_i, y_j)$ quando a malha é refinada. Para simplificar a notação para a equação de diferenças definimos o operador:

$$-\Delta_\delta U_{i,j} = -\left(\frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2} + \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{k^2}\right).$$

Com essa notação as equações discretas (13.75)-(13.76) podem ser reescritas na forma:

$$-\Delta_\delta U_{i,j} = f(x_i, y_j), (x_i, y_j) \in R_\delta \tag{13.77}$$

$$U_{i,j} = g(x_i, y_j), (x_i, y_j) \in \partial R_\delta. \tag{13.78}$$

Substituindo na equação (13.77) cada um dos $(N-1)\times(M-1)$ pontos interiores da malha em R_δ veremos que a função discreta U satisfaz um sistema de equações lineares com $(N-1)\times(M-1)$ equações no mesmo número de incógnitas, incógnitas essas que são as aproximações para a solução da equação diferencial nos pontos da malha. Em notação matricial, seguindo a convenção notacional adotada no caítulo 3, podemos escrever esse sistema como:

$$A\mathbf{U} = \mathbf{c}$$

onde o vetor **U** a matriz A e o vetor **c** são dados respectivamente por:

$$\mathbf{U} = (U_{1,1}, U_{2,1}, \cdots, U_{N-1,1}, U_{1,2}, U_{2,2}, \cdots, U_{N-1,2}, \cdots, U_{1,M-1}, U_{2,M-1}, \cdots, U_{N-1,M-1})^T$$

$$A = \begin{pmatrix} a & b & & & & c & & & \\ b & a & b & & & & \ddots & \mathbf{0} & \\ & \ddots & \ddots & \ddots & & \mathbf{0} & & \ddots & \\ & & \ddots & \ddots & \ddots & & & & c \\ & & & \ddots & \ddots & \ddots & & & \\ c & & & & \ddots & \ddots & \ddots & & \\ & \ddots & & \mathbf{0} & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & \ddots & & & & b & a & b \\ & & & c & & & & b & a \end{pmatrix} \tag{13.79}$$

$$\mathbf{c} = \begin{pmatrix} f(x_1, y_1) + \dfrac{g(x_0, y_1)}{h^2} + \dfrac{g(x_1, y_0)}{k^2} \\ f(x_2, y_1) + \dfrac{g(x_2, y_0)}{k^2} \\ \vdots \\ \\ f(x_{N-1}, y_1) + \dfrac{g(x_N, y_1)}{h^2} + \dfrac{g(x_{N-1}, y_0)}{k^2} \end{pmatrix}$$

Na matriz A os números a, b e c são os coeficientes da discretização de 5 pontos e são dados por:

$$a = \frac{2}{h^2} + \frac{2}{k^2}, \qquad c = -\frac{1}{k^2}.$$

Já o valor de b não é constante na matriz toda. Em algumas posições esse valor é nulo. Essas posições correspondem àqueles pontos situados sobre a fronteira que são dados pelas posições da matriz A $(p*(M-1), p*(M-1)+1)$ e $(p*(M-1)+1, p*(M-1))$ com $p = 1, 2, \ldots$. Nas demais posições o valor de b é $b = -\frac{1}{h^2}$

Talvez essas idéias fiquem mais claras se considerarmos um exemplo.

Exemplo 13.7.1 *Consideremos a malha mostrada na figura abaixo para o domínio* $0 < x < 1$ *e* $0 < y < 1$*.*

Figura 4.2: Malha de discretização com $h = k = \frac{1}{3}$

Existem 4 pontos internos: $P_{1,1}, P_{1,2}, P_{2,1}, P_{2,2}, (P_{i,j} = (x_i, y_j))$. *Nesse caso* $N = M = 3$. *As condições de fronteira para o problema de Dirichlet são dadas como:*

$$u(0, y) = 0, \qquad u(1, y) = 1, \qquad u(x, 0) = 0, \qquad u(x, 1) = 0$$

Observe que estamos utilizando $h = k$ e portanto a equação (13.75) pode ser reescrita como:

$$4U_{i,j} - U_{i+1,j} - U_{i-1,j} - U_{i,j+1} - U_{i,j-1} + h^2 f_{i,j}. \tag{13.80}$$

Variando os índices i e j obtemos:

$$\begin{aligned}
4U_{1,1} - U_{2,1} - U_{0,1} - U_{1,2} - U_{1,0} &= \frac{1}{9} f_{1,1} \\
4U_{1,2} - U_{2,2} - U_{0,2} - U_{1,3} - U_{1,1} &= \frac{1}{9} f_{1,2} \\
4U_{2,1} - U_{3,1} - U_{1,1} - U_{2,2} - U_{2,0} &= \frac{1}{9} f_{2,1} \\
4U_{2,2} - U_{3,2} - U_{1,2} - U_{2,3} - U_{2,1} &= \frac{1}{9} f_{2,2}.
\end{aligned} \tag{13.81}$$

Todos os valores de $U_{i,j}$ para $i = 0$ ou $i = 3$ e qualquer j e $j = 0$ ou $j = 3$ e qualquer i são conhecidos. Substituindo esses valores nas equações (13.81) acima podemos reescrevê-las na forma matricial como:

$$\begin{pmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{pmatrix} \begin{pmatrix} U_{1,1} \\ U_{1,2} \\ U_{2,1} \\ U_{2,2} \end{pmatrix} = \begin{pmatrix} f_{1,1}/9 \\ f_{1,2}/9 \\ 1 + f_{2,1}/9 \\ 1 + f_{2,2}/9 \end{pmatrix}.$$

13.8 Erro de Truncamento Local

Comentamos na seção anterior que a aproximação gerada pela discretização por diferenças finitas "converge" para a solução do problema (13.71-13.72). Nesta seção vamos demonstrar esse fato para o caso de um domínio retangular. Para esse fim precisamos introduzir alguns conceitos e resultados.

Note que a expressão (13.74) não é uma igualdade, de forma que podemos definir a quantidade:

$$T_{i,j} = -\Delta_\delta u_{i,j} - f(x_i, y_j). \tag{13.82}$$

Lema 13.1 *Se a solução $u(x,y)$ é diferenciável até ordem 4 com derivada limitada, então o erro de truncamento local definido por (13.82) satisfaz:*

$$|T_{i,j}| \leq c_1 h^2 + c_2 k^2$$

onde c_1 e c_2 são constantes independentes de h e k.

Prova: Da definição do operador Δ_δ temos o seguinte:

$$\begin{aligned} T_{i,j} = \quad & -\Bigg(\frac{u(x_i+h,y_j) - 2u(x_i,y_j) + u(x_i-h,y_j)}{h^2} \\ & + \frac{u(x_i,y_j+k) - 2u(x_i,y_j) + u(x_i,y_j-k)}{k^2} \Bigg) - f(x_i,y_j). \end{aligned} \tag{13.83}$$

Expandindo os termos $u(x_i+h,y_j), u(x_i-h,y_j), u(x_i,y_j+k)$ e $u(x_i,y_j-k)$ em série de Taylor em torno do ponto (x_i,y_j) obtemos:

$$\begin{aligned} u(x_i+h,y_j) &= u(x_i,y_j) + hu_x(x_i,y_j) + \frac{h^2}{2!}u_{xx}(x_i,y_j) + \frac{h^3}{3!}u_{xxx}(x_i,y_j) + \frac{h^4}{4!}u_{xxxx}(\xi_1,y_j) \\ u(x_i-h,y_j) &= u(x_i,y_j) - hu_x(x_i,y_j) + \frac{h^2}{2!}u_{xx}(x_i,y_j) - \frac{h^3}{3!}u_{xxx}(x_i,y_j) + \frac{h^4}{4!}u_{xxxx}(\xi_2,y_j) \\ u(x_i,y_j+k) &= u(x_i,y_j) + ku_y(x_i,y_j) + \frac{k^2}{2!}u_{yy}(x_i,y_j) + \frac{k^3}{3!}u_{yyy}(x_i,y_j) + \frac{k^4}{4!}u_{yyyy}(x_i,\eta_1) \\ u(x_i,y_j-k) &= u(x_i,y_j) - ku_y(x_i,y_j) + \frac{k^2}{2!}u_{yy}(x_i,y_j) - \frac{k^3}{3!}u_{yyy}(x_i,y_j) + \frac{k^4}{4!}u_{yyyy}(x_i,\eta_2) \end{aligned}$$

onde ξ_1, ξ_2, η_1 e η_2 são pontos arbitrários nos intervalos $\xi_1 \in (x_i, x_i + h)$, $\xi_2 \in (x_i-h,x_i), \eta_1 \in (y_j, y_j+k), \eta_2 \in (y_j-k,y_j)$.

Substituindo essas expansões em (13.83) e simplificando os termos semelhantes obtemos:

$$\begin{aligned} T_{i,j} \quad &= -\Bigg[u_{xx}(x_i,y_j) + \frac{h^2}{4!}u_{xxxx}(\xi_1,y_j) + \frac{h^2}{4!}u_{xxxx}(\xi_2,y_j) \\ & +u_{yy}(x_i,y_j) + \frac{k^2}{4!}u_{yyyy}(x_i,\eta_1) + \frac{k^2}{4!}u_{yyyy}(x_i,\eta_2) - f(x_i,y_j) \Bigg]. \end{aligned} \tag{13.84}$$

Da hipótese que as derivadas u_{xxxx} e u_{yyyy} existem e são limitadas segue que os números $P = max|u_{xxxx}(x,y)|$ e $Q = max|u_{yyyy}(x,y)|$ estão bem definidos. Assim tomando módulo em seguida o máximo da expressão (13.84) e considerando que $u_{xx} + u_{yy} = f$ obtemos:

$$|T_{i,j}| \leq \frac{P}{12}h^2 + \frac{Q}{12}k^2. \tag{13.85}$$

□

Observe que nesse caso o erro de truncamento local decresce ao refinarmos a malha, isto é, ao fazermos h e k menores. Na verdade o lema 13.1 nos diz mais do que simplesmente que o erro diminui, ele nos dá a velocidade de convergência desse erro, que nesse caso é quadrática. No entanto, isto não implica necessáriamente que o erro global definido por;

$$e_{i,j} = u_{i,j} - U_{i,j} \tag{13.86}$$

também está decrescendo. Provaremos esse fato à seguir; mas para esse fim necessitamos de alguns resultados preliminares que passamos a apresentar e demonstrar.

Teorema 13.5 *(a) Se $V(x,y)$ é uma função discreta (de malha) definida sobre $R_\delta \bigcup \partial R_\delta$ e satisfaz*

$$\Delta_\delta V(x,y) \geq 0, \quad \forall (x,y) \in R_\delta,$$

então,

$$\max_{(x,y)\in R_\delta} V(x,y) \quad \leq \quad \max_{(x,y)\in \partial R_\delta} V(x,y).$$

(b) Alternativamente, se $V(x,y)$ satisfaz

$$\Delta_\delta V(x,y) \leq 0, \quad \forall (x,y) \in R_\delta,$$

então,

$$\min_{(x,y)\in R_\delta} V(x,y) \quad \geq \quad \min_{(x,y)\in \partial R_\delta} V(x,y).$$

Prova: Provaremos a parte (a) por contradição. Suponhamos que em algum ponto $P_0 \equiv (x_r, y_s)$ de R_δ temos $V(P_0) = M_0$ onde $M_0 \geq V(P), \forall P \in R_\delta$ e $M_0 > V(P), \forall P \in \partial R_\delta$.

Sejam

$$\begin{aligned} P_1 &= (x_r + h, y_s), \\ P_2 &= (x_r - h, y_s), \\ P_3 &= (x_r, y_s + k) \text{ e} \\ P_4 &= (x_r, y_s + k). \end{aligned}$$

Então, usando (13.75) podemos escrever

$$\Delta_\delta V(P_0) \equiv \frac{V(P_1)+V(P_2)}{h^2} + \frac{V(P_3)+V(P_4)}{k^2} - 2\left(\frac{1}{h^2}+\frac{1}{k^2}\right)V(P_0).$$

Mas, por hipótese, temos $\Delta_\delta V(P_0) \geq 0$, de modo que

$$M_0 = V(P_0) \leq \frac{1}{1/h^2+1/k^2}\left(\frac{1}{h^2}\frac{V(P_1)+V(P_2)}{2} + \frac{1}{k^2}\frac{V(P_3)+V(P_4)}{2}\right). \tag{13.87}$$

Como $M_0 \geq V(Q), \quad \forall Q \in R_\delta \cup \partial R_\delta$, implica que $V(P_\nu) = M_0$ para $\nu = 1,2,3,4$, pois se $V(P_\nu) < M_0$ para algum $\nu = 1,2,3,4$ então (13.87) implica o seguinte:

$$M_0 = V(P_0) < \frac{1}{1/h^2+1/k^2}\left(\frac{1}{h^2}\frac{M_0+M_0}{2} + \frac{1}{k^2}\frac{M_0+M_0}{2}\right) = M_0$$

o que leva a uma contradição pois M_0 não pode ser estritamente menor do que ele mesmo. Lembre-se que estamos supondo que M_0 é o máximo de V em todo o domínio e portanto $V(Q) \leq M_0 \quad \forall Q$.

Agora, repetimos esse argumento para cada um dos pontos interiores P_ν no lugar de P_0. Por repetição, cada ponto de R_δ e ∂R_δ aparece como um dos pontos P_ν para algum correspondente P_0. Assim, concluímos que

$$V(P) = M_0 \text{ para todo } P \in R_\delta \cup \partial R_\delta,$$

o que contradiz a hipótese que $V < M_0$ em ∂R_δ. Daí, a parte (a) do teorema segue.[1]

Para provar a parte (b), podemos repetir o argumento acima. Entretanto, é mais simples recordar que

[1] Na verdade provamos mais que isso. Provamos que se o máximo, no caso (a) ou o mínimo, no caso (b), de $V(x,y)$ ocorre em R_δ, então $V(x,y)$ é constante em R_δ e ∂R_δ.

$$\max[-V(x,y)] = -\min V(x,y),$$
$$\Delta_\delta(-V) = -\Delta_\delta(V).$$

Portanto, se V satisfaz a hipótese da parte (b), então $-V$ satisfaz a hipótese da parte (a). Mas a conclusão da parte (a) para $-V$ é a mesma da parte (b) para V. □

Aplicando adequadamente o princípio do máximo podemos obter um limitante para a solução da equação de diferenças (13.75). O resultado, chamado *estimativa a priori*, pode ser dado por:

Teorema 13.6 *Seja $V(x,y)$ qualquer função discreta definida sobre os conjuntos R_δ e ∂R_δ. Então,*

$$\max_{(x,y)\in R_\delta} |V(x,y)| \leq \max_{(x,y)\in \partial R_\delta} |V(x,y)| + \frac{a^2}{2} \max_{(x,y)\in R_\delta} |\Delta_\delta V(x,y)| \tag{13.88}$$

Prova: Vamos introduzir a função

$$\phi(x,y) \equiv \frac{1}{2}x^2$$

e observemos que para todo $(x,y) \in R_\delta \bigcup \partial R_\delta$,

$$0 \leq \phi(x,y) \leq \frac{a^2}{2} \quad \text{e} \quad \Delta_\delta \phi(x,y) = 1.$$

Agora, sejam $V_+(x,y)$ e $V_-(x,y)$ dadas por

$$V_\pm(x,y) \equiv \pm V(x,y) + N_0\phi(x,y),$$

onde

$$N_0 \equiv \max_{(x,y)\in R_\delta} |\Delta_\delta V(x,y)|.$$

Para todo $(x,y) \in R_\delta$ é fácil mostrar que (ver exercício 13.36)

$$\Delta_\delta V_\pm(x,y) = \pm\Delta_\delta V(x,y) + N_0 \geq 0.$$

Assim, podemos aplicar o princípio do máximo, parte (a) do Teorema 13.5, a cada $V_\pm(x,y)$ e obtemos, para todo $(x,y) \in R_\delta$,

$$V_\pm(x,y) \leq \max_{(x,y)\in \partial R_\delta} V_\pm(x,y) =$$

$$\max_{(x,y)\in \partial R_\delta} [\pm V(x,y) + N_0\phi] \leq \max_{(x,y)\in \partial R_\delta} [\pm V(x,y)] + N_0\frac{a^2}{2}.$$

Mas, da definição de $V_\pm$ e do fato que $\phi \geq 0$, obtemos:

$$\pm V(x,y) \leq V_\pm(x,y).$$

Portanto,

$$\pm V(x,y) \leq \max_{(x,y)\in \partial R_\delta} [\pm V(x,y)] + N_0\frac{a^2}{2} \leq \max_{(x,y)\in \partial R_\delta} |V(x,y)| + \frac{a^2}{2}N_0.$$

Como o lado direito da desigualdade acima é independente de $(x,y) \in R_\delta$ o teorema segue. □

Observação 13.8.1 *Note que podemos substituir* $\frac{a^2}{2}$ *em (13.88) por* $\frac{b^2}{2}$ *desde que usemos* $\psi(x,y) = \frac{y^2}{2}$ *no lugar de* $\phi(x,y)$ *na prova do teorema.*

Para encontrar uma estimativa para o erro global, tomaremos no teorema 13.6, $V(x,y) = e(x,y)$ então segue que:

$$\pm e_{i,j} \leq max_{\partial R_\delta}|e_{i,j}| + \frac{a^2}{2}N_0.$$

Mas como $U_{i,j} = u_{i,j} = g_{i,j}$ na fronteira ∂R_δ temos que $e_{i,j} = 0$ em ∂R_δ. Assim,

$$\pm e_{i,j} \leq \frac{a^2}{2}N_0$$

ou seja,

$$|e_{i,j}| \leq \frac{a^2}{2}N_0 = \frac{a^2}{2}max_{R_\delta}|\Delta_\delta e_{i,j}|.$$

Agora,

$$\Delta_\delta e_{i,j} = \Delta_\delta(u(x_i,y_j) - U_{i,j}) = \Delta_\delta u(x_i,y_j) - \Delta_\delta U_{i,j} = \Delta_\delta u_{i,j} - f_{i,j} = T_{i,j},$$

usando (13.82). Assim,

$$|e_{i,j}| \leq \frac{a^2}{2}max_{R_\delta}|T_{i,j}| \leq \frac{a^2}{24}(Ph^2 + Qk^2).$$

Portanto o método numérico definido em (13.75) produz uma solução $U_{i,j}$ que converge pontualmente para a solução $u(x_i,y_j)$ quando $h \to 0$ e $k \to 0$ com $x_i = ih, y_j = jk$ valores fixos.

Corolário 13.1 *Uma consequência do teorema 13.6 é que o sistema de equações lineares* $A\mathbf{U} = \mathbf{c}$ *definido pela matriz (13.79) tem uma única solução.*

Prova: Provaremos que o sistema homogêneo correspondente a (13.75) tem somente a solução trivial. Para isso consideremos o problema homogêneo:

$$-(u_{xx} + u_{yy}) = 0 \text{ em } R$$

com condição de fronteria homogênea $u = 0$ sobre a fronteira ∂R, cuja única solução é óbviamente a solução $u \equiv 0$. Discretizando esse problema como fizemos para obter (13.75) obtemos um sistema linear homogêneo para as incógnitas $U_{i,j}$. Utilizando agora o teorema 13.6 com $V(x,y) = u(x,y)$, teremos:

$$max_{R_\delta}U_{i,j} \leq max_{\partial R_\delta}U_{i,j} = 0.$$

Por outro lado como $\Delta_\delta U_{i,j} = 0$ pois estamos assumindo que $U_{i,j}$ é solução da equação de diferenças, temos também , usando a parte (b) do teorema 13.5:

$$min_{R_\delta}U_{i,j} \geq min_{\partial R_\delta}U_{i,j} = 0.$$

Assim, $U_{i,j} = 0$ é a unica solução do sistema em questão. □

13.9 Condições de Fronteira em Domínios Gerais

Na prática raramente os problemas apresentam-se em domínios retangulares, de forma que os métodos apresentados na seção anterior teriam pouco valor na solução de problemas reais não fossem eles estendíveis para domínios mais gerais. Nesta seção apresentaremos duas técnicas para aproximação numérica da equação de Poisson com condição de fronteira de Dirichlet em domínios gerais. O caso da condição de Neumann é bem mais complexo e será tratado em outra seção. Consideremos então o problema de Dirichlet num domínio R como o da figura 4.3 abaixo.

Figura 4.3: Domínio Geral - Aproximação por lados da malha

Observação 13.9.1 *Na figura 4.3 e seguintes estaremos utilizando a localização geográfica para rotular os pontos da discretização de 5 pontos, isto é, o ponto central da discretização chamaremos de C, N para o ponto acima (Norte), S para o ponto abaixo (Sul), L para o ponto à direita (Leste) e O para o ponto à esquerda (Oeste).*

Notemos que ao contrário do caso de domínio retangular, neste caso os pontos onde a condição de fronteira é conhecida não fazem parte da malha e portanto não podemos utilizar a condição de fronteira para eliminá-los da equação (13.75). A primeira técnica que apresentamos consiste em aproximar a fronteira do domínio por segmentos da malha como ilustrado na figura 4.3 acima pela linha mais grossa. Completada essa aproximação passamos a resolver o problema no novo domínio supondo que a condição de fronteira $u(x, y) = g(x, y)$ aplica-se agora sobre a nova fronteira. Como pode ser facilmente e diretamente observado da figura 4.3, ao refinarmos a malha melhores aproximações da fronteira são obtidas. Entretanto, esse não é o método mais preciso que podemos deduzir. Sua grande vantagem está na simplicidade e facilidade de aplicação. Na prática, quando aproximamos o domínio pelos lados das células e transportamos a condição de fronteira para essa nova curva, (observe o ponto P na figura 4.3), estamos aproximando o valor de U_O por aquele de U_P, ou seja, estamos interpolando u na célula por um polinômio constante de grau zero. Como é sabido da teoria da aproximação [**?**], o erro em tal aproximação é de ordem h. Já o erro de aproximação da equação diferencial por diferenças finitas, como foi mostrado em (13.85), é de ordem h^2. Por essa razão foi comentado acima ser esse um método que não goza de boa precisão.

O sistema linear resultante da aplicação dessa técnica é similar àquele obtido para o caso de um domínio retangular tendo a forma $A\mathbf{U} = \mathbf{c}$ onde $\mathbf{U}$ representa o vetor das incógnitas, a matriz A tem as mesmas características daquela em (13.79) a menos do fato de que a subdiagonal onde aparece a constante c deixa de ser uma diagonal e os valores aparecem em zig-zag.

Exemplo 13.9.1 *Considere como um exemplo, a equação de Poisson com condição de fronteira de Dirichlet para o domínio da figura 4.4.*

Figura 4.4:Exemplo de discretização com 24 pontos internos.

A discretização dessa equação por diferenças finitas de cinco pontos tomando como aproximação da fronteira irregular os lados da malha, conforme ilustrado pela linha cheia na figura 4.4, obtemos um sistema linear 24×24 $A\mathbf{U} = \mathbf{c}$ *onde:*

$$\mathbf{U} = (U_1, U_2, \ldots, U_{24})^T$$

com U_i *denotando uma aproximação para* $u(x, y)$ *no ponto da figura marcado com* i.

$$A = \left\{ \begin{array}{cccccccccccccccccccccccc}
a & & & c & & & & & & & & & & & & & & & & & & & & \\
& a & b & & & & c & & & & & & & & & & & & & & & & & \\
& b & a & b & & & & c & & & & & & & & & & & & & & & & \\
c & & b & a & b & & & & c & & & & & & & & & & & & & & & \\
& & & b & a & & & & & c & & & & & & & & & & & & & & \\
& & & & & a & b & & & & & c & & & & & & & & & & & & \\
& c & & & & b & a & b & & & & & c & & & & & & & & & & & \\
& & c & & & & b & a & b & & & & & c & & & & & & & & & & \\
& & & c & & & & b & a & b & & & & & c & & & & & & & & & \\
& & & & c & & & & b & a & & & & & & c & & & & & & & & \\
& & & & & & & & & & a & b & & & & & & & & & & & & \\
& & & & & c & & & & & b & a & b & & & & c & & & & & & & \\
& & & & & & c & & & & & b & a & b & & & & c & & & & & & \\
& & & & & & & c & & & & & b & a & b & & & & c & & & & & \\
& & & & & & & & c & & & & & b & a & b & & & & c & & & & \\
& & & & & & & & & c & & & & & b & a & & & & & c & & & \\
& & & & & & & & & & & c & & & & & a & b & & & & c & & \\
& & & & & & & & & & & & c & & & & b & a & b & & & & c & \\
& & & & & & & & & & & & & c & & & & b & a & b & & & & c \\
& & & & & & & & & & & & & & c & & & & b & a & b & & & \\
& & & & & & & & & & & & & & & c & & & & b & a & & & \\
& & & & & & & & & & & & & & & & c & & & & & a & b & \\
& & & & & & & & & & & & & & & & & c & & & & b & a & b \\
& & & & & & & & & & & & & & & & & & c & & & & b & a
\end{array} \right\}$$

onde

$$a = \frac{2}{h^2} + \frac{2}{k^2}, \quad b = -\frac{1}{h^2} \quad e \quad c = -\frac{1}{k^2}$$

O vetor $\mathbf{c}$ *contém o valor da função* $f(x,y)$ *avaliada nos pontos correspondentes à enumeração da figura 4.4 e também os valores da fronteira, por exemplo a primeira coordenada de* $\mathbf{c}$ *é:*

$$c_1 = f(x_5, y_1) + \frac{g(x_4, y_1) + g(x_6, y_1)}{h^2} + \frac{g(x_5, y_0)}{k^2}$$

que corresponde ao valor de f *no ponto* 1 *adicionado aos valores da fronteira correspondentes aos pontos à esquerda e à direita de* 1 *e tamém do ponto abaixo.*

As demais coordenadas de $\mathbf{c}$ *são calculadas de maneira similar e deixamos como exercício. Ver Exercício (13.37).)*

A outra técnica consiste na utilização de um polinômio de primeiro grau na interpolação. No caso da figura 4.3 utilizamos os pontos P e C para interpolação e avaliamos esse polinômio no ponto O, na verdade uma extrapolação! Assim o valor de U_O será expresso como função dos valores de U_P que é conhecido e de U_C que não o é. Isto produz uma equação que permite a eliminação de U_O da equação de diferenças para U_C.

Deduzimos à seguir as fórmulas para o caso especial do ponto C da figura 4.3. Na notação da figura 4.3 temos

$$\frac{U_O - 2U_C + U_L}{h^2} + \frac{U_N - 2U_C + U_S}{k^2} = f_C. \tag{13.89}$$

A grande dificuldade de (13.89) comparado com (13.75) é que em (13.89) o valor de U_O é desconhecido pois este não está sobre a fronteira como seria o caso quando o domínio é retangular. Observe na figura 4.4 a ampliação de uma parte da figura 4.3 onde mostramos uma célula onde a fronteira do domínio corta seus lados.

Figura 4.4: Interpolação Linear

O polinômio linear que interpola U_P e U_C é

$$[(x - x_C)U_P - (x - x_P)U_C]\,\frac{1}{x_P - x_C}.$$

A aproximação para U_O pode então ser facilmente deduzida substituindo x por x_O na expressão acima para obter:

$$\begin{aligned} U_O \simeq [(x_O - x_C)U_P - (x_O - x_P)U_C]\,\frac{1}{x_P - x_C} &= [-hU_P + (1-\theta_1)hU_C]\,\frac{-1}{h\theta_1} \\ &= \frac{1}{\theta_1}\,[U_P - (1-\theta_1)U_C]\,. \end{aligned}$$

Da mesma forma

$$U_S \simeq \frac{1}{\theta_2}\left[U_Q - (1-\theta_2)U_C\right].$$

Substituindo essas aproximações em (13.89) obtemos:

$$\frac{1}{h^2}\left[\frac{U_P-(1-\theta_1)U_B}{\theta_1} - 2U_B + U_C\right] + \frac{1}{k^2}\left[\frac{U_Q-(1-\theta_2)U_B}{\theta_2} - 2U_B + U_E\right] = f_B.$$

Eliminando agora os termos semelhantes e passando para o segundo membro os termos conhecidos obtemos a equação final:

$$\frac{1}{h^2}\left[U_L - \frac{(1+\theta_1)U_C}{\theta_1}\right] + \frac{1}{k^2}\left[U_N - \frac{(1+\theta_2)U_C}{\theta_2}\right] = f_C - \frac{U_P}{h^2\theta_1} - \frac{U_Q}{k^2\theta_2}. \tag{13.90}$$

Uma variação da técnica de interpolação acima descrita e que é muitas vezes preferida na prática por tratar-se de interpolação propriamente dita e não extrapolação é a seguinte:

Aplica-se a equação de diferenças de 5 pontos somente para aqueles pontos da malha interiores ao domínio e para os quais todos os 4 vizinhos estão no domínio. Para pontos interiores ao domínio com algum vizinho fora dele, calculamos uma aproximação por interpolação. Assim, para o ponto C da figura 4.3 calculamos uma primeira aproximação para U_C interpolando na direção x os pontos U_P e U_L em seguida calculamos uma segunda aproximação para U_C interpolando na direção y os pontos U_Q e U_N e finalmente adotamos como aproximação definitiva para U_C a média ponderada pelas distâncias dos dois valores obtidos, formalmente:

Calculamos o polinômio que interpola U_P e U_L dado por:

$$\left[(x-x_P)U_L - (x-x_L)U_P\right]\frac{1}{x_L - x_P}.$$

Avaliamos esse polinômio no ponto x_C para obter a primeira aproximação U_C^1 de U_C.

$$U_C^1 = \frac{h\theta_1 U_L + hU_P}{h + h\theta_1} = \frac{\theta_1 U_L + U_P}{1+\theta_1}.$$

Da mesma forma calculamos uma segunda aproximação U_C^2 de U_C interpolando na outra direção e obtemos:

$$U_C^2 = \frac{h\theta_2 U_N + hU_Q}{h + h\theta_2} = \frac{\theta_2 U_N + U_Q}{1+\theta_2}.$$

Finalmente uma aproximação para U_C pode então ser derivada tomando a média ponderada:

$$U_C = \frac{h\theta_1 U_C^1 + h\theta_2 U_C^2}{h\theta_1 + h\theta_2} = \frac{\theta_1 U_C^1 + \theta_2 U_C^2}{\theta_1 + \theta_2}.$$

Maiores detalhes sobre as técnicas aqui descritas podem ser encontrados em [**?**] e [**?**].

13.10 Condição de Fronteria de Neumann

Consideramos nesta seção o problema de Poisson com condição de fronteira com derivadas, ou seja, condições de fronteira de Neumann. No caso de domínio retangular o problema é:

Figura 4.5: Problema de Neumann no retângulo

Diferentemente do caso da condição de Dirichlet onde o valor de u é conhecido na fronteira, no presente caso devemos determinar u tambem nos pontos da fronteira. De forma que precisamos considerar a equação de diferenças finitas para pontos como o ponto C da figura 4.5. Como a equação de diferenças utiliza pontos para traz e para frente (também para cima e para baixo) teremos que introduzir pontos fantasmas que estão fora do domínio de cálculo, como aqueles formados pelas linhas tracejadas da figura 4.5. Utilizamos então as condições de fronteira para eliminá-los. Assim, a equação de Poisson discretizada no ponto C fica:

$$\frac{U_O - 2U_C + U_L}{h^2} + \frac{U_N - 2U_C + U_S}{k^2} = f_C. \tag{13.91}$$

O ponto O é fantasma e deve ser eliminado. Descrevemos à seguir como eliminar esse ponto. Procede-se de maneira similar na eliminação de pontos sobre as outras linhas tracejadas.

Com esse objetivo aproximamos a condição de fronteira $\frac{\partial u(0,y)}{\partial x} = f_4(y)$ por diferenças centrais no ponto C para obter:

$$f_4(y_C) = \frac{U_L - U_O}{2h}$$

ou seja

$$U_O = U_L - 2hf_4(y_C). \tag{13.92}$$

Portanto (13.91) transforma-se em:

$$\frac{U_L - 2hf_4(y_C) - 2U_C + U_L}{h^2} + \frac{U_N - 2U_C + U_S}{k^2} = f_C.$$

Eliminado os termos comuns e passando aqueles conhecidos para o segundo membro obtemos:

$$\frac{2U_L - 2U_C}{h^2} + \frac{U_N - 2U_C - U_S}{k^2} = f_C + \frac{2f_4(y_C)}{h^2}. \tag{13.93}$$

Observando as equações (13.93) e (13.90) concluimos que o efeito das condições de fronteira no primeiro caso e da irregularidade da fronteira no segundo, sobre as equações discretizadas é a modificação do termo independente no lado direito dessas equações e também a modificação de alguns elementos da matriz de coeficientes. As modificações na matriz são as mais relevantes pois apesar de somente uns poucos elementos sofrerem modificações, estas podem ser suficientes para destruir propriedades importantes da matriz tais como simetria e diagonal dominância.

O caso de domínios irregulares com condição de Neumann é muito mais complicado para o tratamento com diferenças finitas. A grande dificuldade reside no fato de conhecermos a derivada direcional $\frac{\partial u}{\partial \eta}$ sobre a fronteira do domínio de forma que ao, por exemplo, aproximarmos o domínio pelos lados das células da discretização, como descrito para o caso da condição de Dirichlet, não podemos simplesmente transportar essas condições para os pontos da malha pois precisamos levar em consideração os cossenos diretores das derivadas direcionais. Isto torna esse expediente extremamente tedioso e complexo o que leva alguns autores [**?**] a sugerir que esses termos devam ser ignorados e devemos então considerar as condições de fronteira sobre a malha como se aí fosse realmente sua localização. Esse processo obviamente introduz erros que são de dificil análise.

13.11 Diferenças Finitas em Coordenadas Polares

A mudança de variáveis $x = r\cos(\theta), y = r\,\mathrm{sen}\,(\theta)$ transforma a equação de Poisson em:

$$\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2} = h(r,\theta) \tag{13.94}$$

Essa transformação pode ser bastante simplificadora no tratamento da discretização do domínio no caso deste conter formas circulares, pois nesse caso as derivadas em (r,θ) são derivadas direcionais nas direções do raio e do ângulo, de forma que podemos utilizar uma malha como mostrado na figura 4.6.

Figura 4.6: Discretização em coordenadas polares

A equação aproximada no ponto (i,j) da malha toma então a forma:

$$\begin{aligned}\frac{U_{i+1,j} - 2Ui,j + Ui-1,j}{h^2} \quad &+ \quad \frac{1}{ih}\left(\frac{U_{i+1,j} - Ui-1,j}{2h}\right)\\ &+ \quad \frac{1}{(ih)^2}\left(\frac{U_{i,j+1} - 2Ui,j + Ui,j-1}{\delta\theta}\right) = h(ih, j\delta\theta)\end{aligned}$$

Notemos que a equação acima tem a mesma forma da equação (13.75) a menos dos termos resultantes da discretização da derivada de primeira ordem. No entanto esse termo irá agrupar-se com aqueles provenientes das derivadas de segunda ordem e ao final teremos o mesmo tipo de molécula de cálculo de 5 pontos tão bem conhecido. Os coeficientes da discretização serão porém distintos daqueles que aparecem na discretização da equação de Poisson em coordenadas cartesianas, como na equação (13.75).

13.12 Exercícios

13.1 *Mostre pelo método de separação de variáveis que a solução do problema (13.18) é dada pela fórmula (13.19).*

13.2 *Seja a equação*

$$u_t = u_{xx} \quad 0 < x < 1$$

com condições de fronteira $u(0,t) = u(1,t) = 0$, *para* $t > 0$ *e com a condição inicial:*

$$u(x,0) = \begin{cases} 2x, & \text{para } 0 \leq x \leq \frac{1}{2} \\ 2 - 2x, & \text{para } \frac{1}{2} \leq x \leq 1 \end{cases}$$

cuja solução exata é:

$$u(x,t) = \frac{8}{\pi^2} \sum_{n=1}^{\infty} (\,sen\,(\frac{1}{2}n\pi))(\,sen\,(n\pi x)) \exp(-n^2\pi^2 t).$$

Faça um programa de computador para resolver esse problema usando o método explícito com $h = 0.05$ *e*

a. $k = \frac{5}{11}h^2$

b. $k = \frac{5}{9}h^2$.

Faça um gráfico da solução numérica e da solução exata para cada um dos casos acima, plotando U *e* u *contra* x *para vários valores de* t, *por exemplo,* $t = 0.05, 0.1, 0.15, \ldots, 0.5$. *Observe atentamente os gráficos dos dois casos. Note que a solução inicial tem uma descontinuidade na primeira derivada no ponto* $x = 1/2$, *e esta descontinuidade é suavizada imediatamente quando entramos no domínio. Esta é uma propriedade geral das equações parabólicas e elípticas. Obtenha pelo método de separação de variáveis a solução exata desse problema.*

13.3 *Mostre que a equação de diferenças*

$$U_{i,j+1} = U_{i,j} + \sigma(U_{i-1,j} - 2U_{i,j} + U_{i+1,j})$$

admite uma solução separável $U_{i,j} = X_i T_j$. *Determine as equações para* X_i *e* T_j. *Obtenha soluções para essas equações substituindo* $X_i = \exp(i\lambda)$ *e determinando dois valores para* λ, *que chamaremos* λ_1 *e* λ_2 *e a solução* X_i *pode ser então determinada por superposição como:*

$$X_i = C_1 \exp(i\lambda_1) + C_2 \exp(i\lambda_2)$$

com C_1 *e* C_2 *constantes arbitrárias.*

13.4 *Mostre que escolhendo* $\sigma = \frac{1}{6}$ *em (13.17) e levando em consideração a expressão (13.21) obtemos um método que é condicionalmente consistente de ordem 2 com a equação do calor* $u_t = au_{xx}$.

13.5 *Mostre que o método explícito*

$$U_{i,j+1} = \frac{2\sigma}{3}U_{i+1,j} + (1-2\sigma)U_{i,j} + \frac{4\sigma}{3}U_{i-1,j}$$

não é consistente com a equação $u_t = au_{xx}$. *Encontre a equação para a qual esse método é consistente.*

13.6 *Mostre que a matrix* $(A)_{pq} = \exp(I\alpha_q ph),\ p = 0, 1, \ldots N,\ q = 0, 1, \ldots N,$ *onde* $\alpha_q = \frac{q\pi}{l}$, *é não singular.*

Sugestão: *Observe que A é uma matriz de Vandermond.*

13.7 *Considere a equação diferencial*

$$u_t = u_{xx} + bu, \quad t > 0, \ \ x \in \mathbb{R}$$

com $b > 0$. *Analise a convergência do método de diferenças finitas:*

$$U_{i,j+1} = \sigma U_{i-1,j} + (1 - 2\sigma + bk)U_{i,j} + \sigma U_{i+1,j}, j = 0, \pm 1, \ldots$$

13.8 *Considere a equação diferencial*

$$\begin{aligned} u_t &= au_{xx} + bu_x, \quad a > 0, \quad 0 \leq x \leq L, \quad 0 < t < T \\ u(x,0) &= \psi(x), \quad 0 < x < L \\ u(0,t) &= f(t), \quad 0 < t < T \\ u(L,t) &= g(t), \quad 0 < t < T. \end{aligned}$$

Discretizando essa equação pelo método explícito e diferenças centrais para aproximar u_x *obtém-se:*

$$U_{i,j+1} = (\sigma + \frac{bk}{2h})U_{i+1,j} + (1 - 2\sigma)U_{i,j} + (\sigma - \frac{bk}{2h})U_{i-1,j}.$$

Utilizando o critério de von Neumann estude a estabilidade desse método. Compare com a estabilidade do método resultante do seguinte critério para aproximar o termo u_x: *(chamado método "up-wind")*

Aproximamos u_x *por diferenças progressivas se* $b > 0$ *e por diferenças regressivas caso contrário.*

13.9 *Considere o seguinte problema de valor de fronteira de Neumann*

$$\begin{aligned} u_t &= au_{xx}, \quad a > 0, \quad 0 \leq x \leq L, \quad 0 < t < T \\ u(x,0) &= \psi(x), \quad 0 < x < L \\ \frac{\partial u(0,t)}{\partial x} &= \alpha_1(u - u_1), \quad 0 < t < T \\ \frac{\partial u(L,t)}{\partial x} &= -\alpha_2(u - u_2), \quad 0 < t < T \end{aligned}$$

onde α_1, α_2 *constantes positivas.*

1. *Obtenha o método resultante da aproximação das condições de fronteira por diferenças centrais e da equação pelo método explícito.*

2. *Obtenha o método resultante da aproximação da fronteira esquerda por diferenças progressivas da fronteira direita por diferenças regressivas e da equação pelo método explícito.*

3. *Repetir os items acima considerando o método implícito para aproximar a equação.*

Escreva esses métodos em notação matricial. Qual o efeito das condições de fronteira na matriz dos coeficientes? Usando o segundo teorema de Gershgorin estude a estabilidade do método resultante do primeiro item, mostrando que a restrição de estabilidade é dada por:

$$\sigma \leq \min\{\frac{1}{2 + \alpha_1 h}, \frac{1}{2 + \alpha_2 h}\}.$$

13.10 *Mostre que se para a equação de diferenças $\mathbf{U}_{j+1} = A\mathbf{U}_j + \mathbf{c}_j$ os autovetores de A são linearmente independentes e se seus autovalores λ_i satisfazem $|\lambda_i| \leq 1 + O(k)$, $\forall i$, então essa equação será estável.*

13.11 *Seja A a matriz tridiagonal de ordem N*

$$A = \begin{pmatrix} a & b & & & & \\ c & a & b & & & \\ & c & a & b & & \\ & & & \ddots & & \\ & & & c & a & b \\ & & & & c & a \end{pmatrix}.$$

Mostre que os autovalores e autovetores dessa matriz são: ($i = 1, 2, \ldots, N$)

$$\lambda_i = a + 2b\left(\frac{c}{b}\right)^{\frac{1}{2}} \cos(\frac{i\pi}{N+1}),$$

$$\mathbf{v}_i = \left((\frac{c}{b})^{\frac{1}{2}}\, sen(\frac{i\pi}{N+1}), (\frac{c}{b})^{\frac{2}{2}}\, sen(\frac{2i\pi}{N+1}), (\frac{c}{b})^{\frac{3}{2}}\, sen(\frac{3i\pi}{N+1}), \cdots, (\frac{c}{b})^{\frac{N}{2}}\, sen(\frac{Ni\pi}{N+1}) \right)^T.$$

*Sugestão: Ver [**?**] páginas 113-115.*

13.12 *Mostre que os autovalores e autovetores da matriz (13.30) são dados por:*

$$\lambda_i = 1 - 4\sigma\, sen^2(\frac{i\pi}{2N}), \quad \mathbf{v}_i = \left(sen(\frac{i\pi}{2N}),\ sen(\frac{2i\pi}{2N}), \cdots,\ sen(\frac{(N-1)i\pi}{2N}) \right)^T.$$

13.13 *Mostre que os autovalores da matriz do método implícito (13.37) são:*

$$\lambda_i = 1 + 4\sigma\, sen^2(\frac{i\pi}{2N})$$

e portanto os autovalores de A^{-1} são $\frac{1}{\lambda_i}$ que são menores que 1 para todo valor de σ implicando que esse método é incondicionalmente estável, como já haviamos concluído.

13.14 *Prove a igualdade:*

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{h^2}\left(\delta_x^2 - \frac{1}{12}\delta_x^4 + \frac{1}{90}\delta_x^6 + \ldots\right) u.$$

Utilize-a para obter a seguinte discretização da equação do calor:

$$\begin{aligned} \frac{U_{i,j+1} - U_{i,j}}{k} &= \frac{1}{2}\left\{ \left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j+1} + \left(\frac{\partial^2 u}{\partial y^2}\right)_{i,j} \right\} \\ &= \frac{1}{2h^2}\left(\delta_x^2 - \frac{1}{12}\delta_x^4 + \frac{1}{90}\delta_x^6 + \ldots\right)(U_{i,j+1} + U_{i,j}) \end{aligned} \tag{13.95}$$

Aplicando o operador $(1 + \frac{1}{12}\delta_x^2)$ a ambos os membros de (13.95) obtemos um método numérico de quarta ordem no espaço:

$$\begin{aligned} (1-6\sigma)U_{i-1,j+1} + (10+12\sigma)U_{i,j+1} &+ (1-6\sigma)U_{i+1,j+1} = \\ &= (1+6\sigma)U_{i-1,j} + (10-12\sigma)U_{i,j} + (1+6\sigma)U_{i+1,j} \end{aligned}$$

Escreva o método acima na forma matricial e utilize o critério da matriz para estudar sua estabilidade.

13.15 *Sejam* $p_1(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$ *e* $p_2(x) = b_l x^l + b_{l-1}x^{l-1} + \cdots + b_1 x + b_0$ *dois polinômios de graus* m *e* l *e* A *uma matriz* $N \times N$*. Se* λ *é um autovalor de* A *com autovetor* v *e se* $p_1(A)$ *for uma matriz inversível então* $\frac{p_2(\lambda)}{p_1(\lambda)}$ *é autovalor de* $(p_1(A))^{-1}p_2(A)$ *com autovetor* v*. Utilize esse resultado para mostrar que os autovalores da matriz de amplificação do método de Crank-Nicolson são:*

$$\lambda_i = \frac{1 - 2\sigma\ sen^2(\frac{i\pi}{2N})}{1 + 2\sigma\ sen^2(\frac{i\pi}{2N})} \quad i = 1, 2, \ldots, N-1.$$

13.16 *Considere a seguinte aproximação para a equação do calor no intervalo* $[0, 1]$*:*

$$\frac{3}{2k}\Delta_t U_{i,j} - \frac{1}{2k}\nabla_t U_{i,j} = \frac{1}{h^2}\delta_x^2 U_{i,j+1}$$

Utilizando o critério da matriz analise a estabilidade desse método.

13.17 *Considere a equação parabólica não linear*

$$u_t = u_{xx}^m, \quad m \quad \text{inteiro } \geq 2$$

que pode ser aproximada pelo método implícito,

$$\frac{U_{i,j+1} - U_{i,j}}{k} = \frac{\theta\delta_x^2(U_{i,j+1})^m + (1-\theta)\delta_x^2(U_{i,j})^m}{h^2}.$$

Expansão em série de Taylor de $(u(x_i, t_{j+1}))^m$*, em torno do ponto* (x_i, t_j) *produz:*

$$\begin{aligned}(u_{i,j+1})^m &= (u_{i,j})^m + k\frac{\partial(u_{i,j})^m}{\partial t} + \cdots \\ &= (u_{i,j})^m + km(u_{i,j})^{m-1}\frac{\partial u_{i,j}}{\partial t} + \cdots .\end{aligned}$$

Dessa forma a menos de termos de ordem maior que $O(k)$*, temos a seguinte aproximação:*

$$(U_{i,j+1})^m = (U_{i.j})^m + m(U_{i,j})^{m-1}(U_{i,j+1} - U_{i,j})$$

que se substituída na equação do método o torna linear. Escrevendo $\omega_i = U_{i+1,j} - U_{i,j}$ *deduza um sistema linear em termos dessa nova variável.*

13.18 *Considere a equação parabólica não linear:*

$$\begin{aligned} u_t &= \phi(x, t, u, u_x, u_{xx}), \quad 0 < x < 1, \ \ 0 < t \leq T, \\ u(x, 0) &= \psi(x), \\ u(0, t) &= f(t), \\ u(1, t) &= g(t). \end{aligned} \tag{13.96}$$

Esse problema constitue um problema bem posto se a condição $\frac{\partial\phi}{\partial u_{xx}} \geq a > 0$ *estiver satisfeita no domínio. Expandindo* $u(x_i, t_{j+1})$ *em série de Taylor no ponto* (x_i, t_j)*, e utilizando a equação (13.96) para substituir* u_t *deduza um método explícito para resolvê-la.*

Uma outra classe importante de equações parabólicas não lineares é:

$$\begin{aligned} u_{xx} &= \phi(x, t, u, u_x, u_t), \quad 0 < x < 1, \ \ 0 < t \leq T, \\ u(x, 0) &= \psi(x), \\ u(0, t) &= f(t), \\ u(1, t) &= g(t). \end{aligned} \tag{13.97}$$

Esse problema constitue um problema bem posto se a condição $\frac{\partial \phi}{\partial u_t} \geq a > 0$ *estiver satisfeita no domínio. Discretizando as derivadas espaciais por diferenças centrais e a derivada temporal por diferença regressiva, no ponto* (x_i, t_{j+1})*, obtenha um método impícito para resolver (13.97). Qual a ordem desse método?*

Considere a equação parabólica quase-linear com

$$\phi(x, t, u_x, u_t) = a(x, t, u)u_x + b(x, t, u)u_t + c(x, t, u).$$

baseando-se no exercício anterior obtenha um método implícito para resolver esse problema que seja linear, isto é cuja solução requeira a solução de um sistema linear apenas. Derive o método de Crank-Nicolson para a equação (13.97).

13.19 *Seja* ψ *a solução da equação do calor* $\psi_t = \psi_{xx}$*. Mostre que se* $\psi = \exp(-\frac{v}{2})$ *e* $u = v_x$*, então* u *é solução da equação não linear:*

$$u_t = u_{xx} - uu_x.$$

Utilizando a transformação acima encontre a solução exata do problema:

$$\begin{aligned} u_t &= u_{xx} - uu_x \quad 0 < x < 1, \;\; t > 0 \\ u(x, 0) &= sen\, \pi x \\ u(0, t) &= u(1, t) = 0 \end{aligned}$$

Resolva esse problema numericamente utilizando um dos métodos discutidos acima e compare seus resultados com a solução exata.

13.20 *Determinar o erro de truncamento local e a estabilidade do método de Hopscotch.*

13.21 *Mostre que cada uma das fórmulas de Saul'yev (**??**-**??**) é incondicionalmente estável.*

13.22 *Utilizando o método da matriz mostre a que condição de estabilidade da equação de diferenças:*

$$\begin{aligned} U_{i,j+1} &= U_{i,j} + \sigma(U_{i-1,j} - 2U_{i,j} + U_{i+1,j}), \; i = 0, 1, \ldots N-1 \\ U_{-1,j} &= U_{1,j} + 20hU_{0,j} \\ U_{N,j} &= 0 \end{aligned}$$

é $\sigma \leq \frac{1}{2+10h}$*. Por outro lado usando o critério de von Neumann obtemos a condição* $\sigma \leq \frac{1}{2}$*. Conclua então que o critério de von Neumann é uma condição necessária mas não suficiente para estabilidade da equação de diferenças.*

13.23 *Mostre que a complexidade algorítmica (número de operações aritméticas) do método explícito para resolver a equação do calor em 2D é: 4 adições + 3 multiplicações por ponto da malha. Mostre também que no caso de um método ADI esse número é: 10 adições + 8 multiplicações + 6 divisões.*

13.24 *Considere o método obtido pela média ponderada dos métodos implícito e explícito:*

$$U_{i,j+1} - U_{i,j} = \sigma\left(\theta \delta_x^2 U_{i,j+1} + (1-\theta)\delta_x^2 U_{i,j}\right), \quad \sigma = \frac{k}{h^2}, \quad 0 \leq \theta \leq 1.$$

Observe que para $\theta = 0$ *obtemos o método explícito, para* $\theta = 1$ *obtemos o método implícito e para* $\theta = 1/2$ *obtemos o método de Crank-Nicolson. Deduzir a expressão do erro de truncamento local e usando a técnica de von Neumann mostre que para* $0 \leq \theta < 1/2$ *o método é estável se* $\sigma \leq \frac{1}{2(1-2\theta)}$ *e incondicionalmente estável para* $1/2 \leq \theta \leq 1$*.*

13.25 *Mostre que os métodos implícito e ADI em 2D são incondicionalmente estáveis.*

13.26 *Mostre que os operadores δ_x^2 e δ_y^2 comutam quando o domínio onde eles se aplicam é um retângulo.Dê um contra-exemplo para ilustrar o caso em que o domínio não é um retângulo.*

13.27 *Mostre que o método de Crank-Nicolson em duas dimensões com espaçamento k na direção t, h_x e h_y nas direções x e y, pode ser escrito como:*

$$\left(1-\frac{\sigma_x}{2}\delta_x^2-\frac{\sigma_y}{2}\delta_y^2\right)U^{n+1}=\left(1+\frac{\sigma_x}{2}\delta_x^2+\frac{\sigma_y}{2}\delta_y^2\right)U^n \qquad (13.98)$$

onde $\sigma_x=\frac{k}{h_x^2}$ e $\sigma_y=\frac{k}{h_y^2}$.

Mostre que o termo:

$$\frac{\sigma_x\sigma_y}{4}\delta_x^2\delta_y^2\left(U^{n+1}-U^n\right)$$

é de ordem $O(k^3)$.

Mostre que adicionando o termo

$$\frac{\sigma_x\sigma_y}{4}\delta_x^2\delta_y^2U^{n+1}$$

no lado esquerdo de (13.98) e o termo

$$\frac{\sigma_x\sigma_y}{4}\delta_x^2\delta_y^2U^{n}$$

no lado direito, obtemos um método com a mesma ordem do método de Crank-Nicolson que pode ser fatorado na forma:

$$\left(1-\frac{\sigma_x}{2}\delta_x^2\right)\left(1-\frac{\sigma_y}{2}\delta_y^2\right)U^{n+1}=\left(1+\frac{\sigma_x}{2}\delta_x^2\right)\left(1+\frac{\sigma_y}{2}\delta_y^2\right)U^n$$

13.28 *Mostre que o método de Peaceman-Rachford para solução da equação do calor não homogênea,*

$$u_t=u_{xx}+u_{yy}+F(x,y,t)$$

toma a forma:

$$\begin{aligned}\left(1-\frac{\sigma_x}{2}\delta_x^2\right)U^{n+\frac{1}{2}} &= \left(1+\frac{\sigma_y}{2}\delta_y^2\right)U^n+\frac{k}{2}F^n\\ \left(1-\frac{\sigma_y}{2}\delta_y^2\right)U^{n+1} &= \left(1+\frac{\sigma_x}{2}\delta_x^2\right)U^{n+\frac{1}{2}}+\frac{k}{2}F^{n+1}\end{aligned}$$

13.29 *Escreva um programa MATLAB para resolver o problema:*

$$\frac{\partial^2u}{\partial x^2}+\frac{\partial^2u}{\partial y^2}=-2$$

definido no quadrado $[-1,1]\times[-1,1]$ com condição de Dirichlet nula na fronteira. Utilize $h=0.2$. Compare seu resultado com a solução exata:

$$u(x,y)=1-y^2-\frac{32}{\pi^3}\sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)^3}\operatorname{sech}\left(\frac{(2n+1)\pi}{2}\right)\cosh\left(\frac{(2n+1)\pi x}{2}\right)\cos\left(\frac{(2n+1)\pi y}{2}\right).$$

13.30 *Escreva um programa MATLAB para resolver o problema:*

$$\frac{\partial^2u}{\partial x^2}+\frac{\partial^2u}{\partial y^2}-10u=0$$

definido no quadrado $[-1,1]\times[-1,1]$ com condições de fronteira dadas por:

1. $u = 0$ *para* $y = 1$ *e* $-1 \le x \le 1$;
2. $u = 1$ *para* $y = -1$ *e* $-1 \le x \le 1$;
3. $\frac{\partial u}{\partial x} = -0.5u$ *para* $x = 1$ *e* $-1 < y < 1$;
4. $\frac{\partial u}{\partial x} = 0.5u$ *para* $x = -1$ *e* $-1 < y < 1$.

Utilize uma malha uniforme com $h = 0.1$.

13.31 *Considere o problema*

$$u_{xx} + u_{yy} = 0$$

definido no domínio:

$$R = \{(x, y), \quad x^2 + y^2 \le 1, \quad x \ge 0, \quad y \ge 0\}$$

que corresponde ao semi-círculo superior de centro na origem e raio 1, com condições de fronteira:

$$u(x, 0) = 0, \quad -1 \le x \le 1, \qquad u(x, y) = 100, \quad x^2 + y^2 = 1,$$

cuja solução analítica é:

$$u(r, \theta) = \frac{400}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \left(\frac{r}{2}\right)^{2n-1} sen\,(2n-1)\theta.$$

Obter a discretização desse problema em coordenadas polares com $\delta_r = 0.1$, $\delta_\theta = \pi/10$. *Usar MATLAB para resolver o sistema linear resultante.*

13.32 *A fórmula de 5 pontos é de ordem* $O(h^2)$. *Se porventura necessitarmos de uma fórmula mais precisa para discretizar a equação de Laplace, devemos utilizar mais pontos. Mostre que uma fórmula de ordem* $O(h^4)$ *pode ser obtida utilizando-se os pontos da molécula abaixo:*

e ela é dada por:

$$U_0 = \frac{1}{60} \left(-U_7 + 16U_3 + 16U_1 - U_5 - U_6 + U_2 + 16U_4 - U_8\right)$$

No entanto essa fórmula não pode ser aplicada para pontos próximos à fronteira, pois ela utiliza dois pontos em cada direção. Poder-se-ia utilizar a molécula de cálculo:

para deduzir o esquema numérico, que também é de ordem $O(h^4)$.

$$U_0 = \frac{1}{20}\left(4(U_1 + U_2 + U_3 + U_4) + U_5 + U_6 + U_7 + U_8\right). \tag{13.99}$$

Mostre que no caso da equação de Poisson o esquema (13.99) transforma-se em:

$$4\left(U_1 + U_2 + U_3 + U_4\right) + U_5 + U_6 + U_7 + U_8 - 20U_0 = -\frac{h^2}{2}\left(8f_0 + f_1 + f_2 + f_3 + f_4\right).$$

Calcule o ETL em todos os casos.

13.33 *Considere a numeração "red-black" como na figura abaixo:*

Figura 4.9: Numeração "red-black"

Obtenha a matriz do sistema linear resultante. Generalize esse resultado para uma malha com N pontos na direção x e M pontos na direção y.

13.34 *A equação biharmônica é dada por:*

$$\Delta\Delta u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = f(x,y).$$

Aproximando o operador $\Delta\Delta u$ por:

$$\Delta\Delta u \approx \alpha_0 U_0 + \alpha_1 \sum_{i=1}^{4} U_i + \alpha_2 \sum_{i=5}^{8} U_i + \alpha_3 \sum_{i=9}^{12} U_i$$

numa malha uniformemente espaçada de h, obtenha valores para os parâmetros $\alpha_0, \alpha_1, \alpha_2$ e α_3 tal que a fórmula tenha ordem $O(h^2)$. Deduza a fórmula do ETL. A molécula computacional é dada na figura:

Figura 4.9: Molécula de cálculo para a equação biharmônica.

13.35 *Considere o problema elíptico:*

$$\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(b(x,y)\frac{\partial u}{\partial y}\right) = f(x,y)$$

com condição de Dirichlet sobre a fronteira de um retângulo. Mostre que a matriz resultante da discretização de 5 pontos é simétrica.

13.36 *Mostre que*

$$\Delta_\delta V_\pm(x,y) = \pm\Delta_\delta V(x,y) + N_0 \geq 0.$$

13.37 *Obtenha os demais elementos do vetor* **c** *do exemplo (13.9.1).*

13.38 *Mostre que a discretização de 5 pontos da equação de Poisson para a molécula da figura 4.9 é dada por:*

$$\alpha_0 U_0 + \alpha_1 U_1 + \alpha_2 U_2 + \alpha_3 U_3 + \alpha_4 U_4 = f_0$$

onde os coeficientes α são dados por:

$$\alpha_0 = -2\left(\frac{1}{h_1 h_3} + \frac{1}{h_2 h_4}\right); \quad \alpha_1 = \frac{2}{h_1(h_1+h_3)}; \quad \alpha_2 = \frac{2}{h_2(h_2+h_4)};$$

$$\alpha_3 = \frac{2}{h_3(h_1+h_3)}; \quad \alpha_4 = \frac{2}{h_4(h_2+h_4)};$$

Figura 4.9: Molécula de Cálculo para a discretização de 5 pontos em pontos não igualmente espaçados.

Calcule o erro de truncamento local e determine a ordem de consistência dessa discretização. Mostre que quando $h_1 = h_3 = h$ *e* $h_2 = h_4 = k$ *essa discretização é de ordem* $O(h^2 + k^2)$*. Explique como essa molécula pode ser utilizada para discretizar a equação de Poisson num domínio geral.*

13.39 *Utilizar a técnica derivada no exercício anterior para aproximar a solução do problema do exercício (13.31) na malha da figura abaixo.*

Utilizar MATLAB na solução do sistema linear.

13.40 *Considere o caso* 4×4 *da matriz* A_h *de (**??**). Utilizando MATLAB calcule os autovalores de* A_h*. Resolva agora o sistema linear resultante pelo método de Jacobi, também utilizando MATLAB e confirme que a taxa de convergência desse método é cerca de* 0.5*. Repita esse mesmo exercício para Gauss-Seidel.*

Capítulo 14

Exercícios Mistos

14.1 - *Mostre que o método de Jacobi diverge para todos os valores de a que tornam a matriz a*

$$A = \begin{pmatrix} 20 & 3 & 1 \\ a & 20 & 1 \\ 1 & a & 6 \end{pmatrix} \; .$$

positiva definida .

14.2 - *Dada uma função f(x), deseja -se calcular:*

$$\int_a^b f(x)dx.$$

Para isso foram propostos dois métodos:

a) *interpolar f(x) em n+1 pontos por um polinômio de grau n e integrá -lo,*

b) *aproximar f(x), usando o método dos mínimos quadrados, por uma função F(x) e integrá - la.*

Usar estes dois métodos para estimar:

$$\int_0^1 \frac{x}{x^2+3x+2}dx \; ,$$

com $n = 4$ *e* $F(x) = a_0 + a_1 x + a_2 x^2$. *Compare seus resultados com o valor exato que é* $ln\frac{9}{8}$.

14.3 - *Para a duplicação de uma avenida, um estudo de engenharia de transportes necessita do cálculo do número total de veículos que passam pela avenida em um período de 24 horas. Um engenheiro vai ao local várias vezes durante um período de 24 horas, e conta o número de carros, por minuto, que passam pela avenida. Os dados obtidos pelo engenheiro encontram-se na tabela a seguir:*

Hora	*Carros/min*	*Hora*	*Carros/min*
0:00	*2*	*12:30*	*15*
2:00	*2*	*14:00*	*7*
4:00	*0*	*16:00*	*9*
5:00	*2*	*17:00*	*20*
6:00	*5*	*18:00*	*22*
7:00	*8*	*19:00*	*10*
8:00	*20*	*20:00*	*11*
9:00	*12*	*21:00*	*8*
10:30	*5*	*22:00*	*5*
11:30	*10*	*23:00*	*5*
		0:00	*3*

a) *Usando interpolação linear estime quantos carros, por minuto, passaram pela avenida às 10:15 horas.*

b) *Usando os dados acima e uma fórmula de quadratura adequada, estime o número de carros que passam pela avenida por dia. (Cuidado com as unidades!).*

Referências Bibliográficas

[1] ACTON, F.S. - Numerical Methods that Work. Harper & Row, Publishers, 1970.

[2] ANTON, H. - Álgebra Linear. Editora Campus, 1982.

[3] ALBRECHT, P. - Análise Numérica: Um curso moderno. LTC, Rio de janeiro, 1973.

[4] BAJPAI, A.C.; MUSTOE, L.R.; WALKER, D. - Matemática para Engenharia. Hmus Livraria Editora Limitada, 1980.

[5] BARNETT, S. - Matrices-Methods and Applications. Clarendon Press. Oxford, 1990.

[6] BARROS, I.Q. - Introdução ao Cálculo Numérico.

[7] BARROS, I.Q. - Métodos Numéricos I - Álgebra Linear.

[8] BRONSON, R., Moderna Introdução às Equações Diferenciais, (Coleção Schaum).

[9] CALLIOLI, C.A.; DOMINGUES, H.H.; COSTA, R.C.F. - Álgebra Linear e Aplicações.Atual Editora Ltda, 1978.

[10] CARNAHAN, B.; LUTHER, H.A. and WILKES, J.O. - Applied Numerical Methods. John Wiley and Sons, Inc; 1969.

[11] COHEN, A.M. et al. - Numerical Analysys. Mc Graw-Hill Book Company (UK) Limited, 1973.

[12] CONTE, S.D. - Elementary Numerical Analysis.

[13] CONTE, S.D.; de Boor, C. - Elementary Numerical Analysis - An Algoritmic Approach, McGraw-Hill International Editions, 1981.

[14] DEMIDOVICH, B.P. and MARON, I.A. - Computational Mathematics.

[15] FORSYTHE, G.E. - Computer Solution of Linear Algebraic Sistems.

[16] FROBERG, W. - Introduction to Numerical Analysis.

[17] GELFAND, I.M. - Lectures on Linear Algebra.

[18] GERALD, C.F., WHEATLEY, P.O.- Applied Numerical Analysis. Addison-Wesley Plublishing Company, 1984.

[19] GOURLAY, A.R.;WATSON,G.A.- Computational Methods for Matrix Eigenproblems. John Wiley & Sons, 1973.

[20] HILDEBRAND, F.B. - Introduction to Numerical Analysis.

[21] HENRICE, P. - Discrete Variable Methods in Ordinary Differential Equations, 1962.

[22] HENRICE, P. - Elements of Numerical Analysis, John Wiley, 1964.

[23] HENRICE, P. - Essentials of Numerical Analysis with Pocket Calculator Demonstrations. John Wiley & Sons, Inc., 1982.

[24] HENRICE, P. - Applied Methods in Computational Complex Analysis, vols I and II. Wiley, New York, 1977

[25] JACQUES, I.; JUDD, C. - Numerical Analysis. Chapman and Hall, Ltd., 1987.

[26] JENNINGS, W. - First Course in Numerical Methods.

[27] KELLY, L.J. - Handbook of Numerical Methods and Applications. Addison- Wesley Publishing Company, 1967.

[28] KNOPP, P.J. - An Introduction Linear Algebra.

[29] LAMBERT, J.D., - Computational Methods in Ordinary Differential Equations , John Wiley and Sons.

[30] LINHARES, O.D., - Cálculo Numérico B. Apostila publicada pelo Departamento de Ciências de Computação e Estatítica do ICMSC, 1969.

[31] LIPCHUTZ, S. - Algebra Linear (Coleção Schaum).

[32] OSTROWSKI, A. M. - Solution of Equations and Systems of Equations. Academic Press, Inc., 1966.

[33] RALSTON, A. - A First Course in Numerical Analysis.

[34] SCHEID, F. - Numerical Analysis (Coleção Schaum).

[35] SCHWARZ, H.R., RUTISHAUSER, H. STIEFEL, E.- Numerical Analysis of Symmetric Matrices. Prentice-Hall, Inc., 1973.

[36] STROUD, A.H., SECREST, D. - Gaussian Quadrature Formulas.

[37] SWOKOWSKI, E. W. - Cálculo com Geometria Analítica. McGraw-Hill do Brasil, Ltda, 1983.

[38] WATSON, W.A., PHILIPSON, T., OATES, P.J. - Numerical Analysis. Edward Arnold (Publishers) Ltd, 1981.

[39] WILKINSON, J.H. - The Algebraic Eigenvalue Problem. Clarendon Press-Oxford, 1965.

[40] YOUNG. D.M. and GREGORY, R.T. - A Survey of Numerical Mathematics.